20th INTERNATIONAL CONFERENCE

ON SCIENTOMETRICS & INFORMETRICS

ISSI 2025

23–27 June 2025

Yerevan, Armenia

PROCEEDINGS

Editors

Shushanik Sargsyan, Wolfgang Glänzel, Giovanni Abramo

UDC 001.8(082)

Sponsors

Diamond sponsor – Higher Education and Science Committee, RA MoESCS

Gold sponsor - Journal of Data and Information Science

Silver sponsor – Clarivate

Bronze sponsor – Quantitative Science Studies

Bronze sponsor – Springer

Partners

Institute for Informatics and Automation Problems of NAS RA

Yerevan State University

Center for Scientific Information Analysis and Monitoring

ISBN 978-9939-1-2086-7

ISSN 2175-1935

© Authors. No part of this book may be reproduced in any form without the written permission of the authors

© International Society for Scientometrics and Informetrics (I.S.S.I.)

© Institute for Informatics and Automation Problems of NAS RA

June 2025 Printed in Armenia

Organizing Committee

Local Members

Shushanik Sargsyan, Armenia, *Chair*

Edita Gzoyan, Armenia

Aram Mirzoyan, Armenia

Gevorg Kesoyan, Armenia

Simon Hunanyan, Armenia

Yeranuhi Manukyan, Armenia

International Members

Jacqueline Leta, Brazil

Vivek Kumar Singh, India

Program Committee Chairs

Sargis Hayotsyan, Armenia

Hovhannes Hovhannisyan, Armenia

Hrachya Astsatryan, Armenia

Giovanni Abramo, Italy

Wolfgang Glanzel, Belgium

Doctoral Forum Committee

Iana Atanassova, France

Andrea Scharnhorst, Netherlands

Gunnar Sivertsen, Norway

Workshops & Tutorials Committee

Cinzia Daraio, Italy

Alesia Zuccala, Denmark

Poster Session Committee

Jacqueline Leta, Brazil

Pei-Shan Chi, Taiwan, Belgium

Eugene Garfield Award Committee

Guillaume Cabanac, France

Nees Jan van Eck, Netherlands

Mike Thelwall, United Kingdom

Lin Zhang, China, Belgium

Student Travel Award Committee

Dag W. Aksnes, Norway

Rodrico Costas, Netherlands

Juan Gorraiz, Austria

Best Paper Award Committee

Nicolas Robinson Garcia, Spain

Vivek Singh, India

Cassidy Sugimoto, USA

Scientific Committee

Dag W. Aksnes, Norway

Iana Atanassova, France

Alberto Baccini, Italy

Rafayel Barkhudaryan, Armenia

Aparna Basu, India

Marc Bertin, France

Sujit Bhattacharya, India

Kevin Boyack, USA

Guillaume Cabanac, France

Zaida Chinchilla-Rodríguez, Spain

Ciriaco Andrea D'Angelo, Italy

Cinzia Daraio, Italy

Gemma Derrick, United Kingdom

Sergio Luiz Monteiro Salles Filho, Brazil

Wolfgang Glänzel, Belgium

Claudia Gonzalez-Brambila, Mexico

Maria Cláudia Cabrini Gracio, Brazil

Edita Gzoyan, Armenia

Robin Haunschild, Germany

Hamid R. Jamali, Australia

Vincent Larivière, Canada

Jacqueline Leta, Brazil

Domenico Augusto Maisano, Italy

Philipp Mayr, Germany

Rogério Mugnaini, Brazil

Bosire Onyancha, South Africa

Gangan Prathap, India

Emanuela Reale, Italy

Ronald Rousseau, Belgium

Shushanik Sargsyan, Armenia

Vivek Kumar Singh, India

Gunnar Sivertsen, Norway

Cassidy Sugimoto, USA

Mike Thelwall, United Kingdom

Lin Zhang, China, Belgium

Yi Zhang, Australia

Alesia Zuccala, Denmark

Design: Anna Margaryan

Layout: Miranush Kesoyan, Mariam Yeghikyan

Preface

It is our great pleasure to present the *Proceedings of the 20th International Conference on Scientometrics & Informetrics (ISSI2025)* of the International Society for Scientometrics and Informetrics, held from June 23 to 27, 2025, in Yerevan, Armenia. This edition of the ISSI conference marks four decades of global exchange and collaboration in the field of scientometrics and informetrics — a field that continues to grow in relevance as science itself evolves in complexity, scope, and global impact.

Hosted for the first time in the South Caucasus, ISSI2025 brought together over 230 participants from more than 38 countries, making it one of the most geographically and thematically inclusive gatherings in the history of the ISSI community. The conference's theme — "Shaping the Future: New Horizons in the Science of Science" — inspired reflection on our field's legacy while encouraging the exploration of bold new directions for scientometric research.

The conference opened with warm welcomes from local and international leaders, including representatives of Armenia's academic and governmental institutions and the President of ISSI.

Two keynote addresses, delivered by **Mike Thelwall** and **Gunnar Sivertsen**, focused on some of the field's most burning questions — including the use of Large Language Models in evaluative contexts and the core values guiding our work in research assessment and policy-relevant application.

During five days, ISSI2025 featured:

- **More than 30 parallel sessions,** showcasing cutting-edge work in areas such as advanced informetric models, science policy and research evaluation, artificial intelligence and scientometrics, open science, micro- and macro-level analysis, technology and innovation studies, and gender, collaboration, and mobility in science.
- **The Doctoral Forum**, providing early-stage researchers an opportunity to present and discuss their work with peers and senior experts in the field.
- **Workshops and Tutorials**, including:
    - The Joint Workshop of the 5th AI + Informetrics (AII) and the 6th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE) — AII-EEKE 2025
    - The tutorial "Exploring the OpenAIRE Graph on Google Big Query", offering hands-on insights into open scholarly data
- **Two special tracks:**
    - *FRAME* (*Framework for the Responsible Use of Assessments and Metrics in Evaluation*), dedicated to developing fair, inclusive, and context-sensitive approaches to research evaluation

    - *Open Research Information (ORI)*, focused on infrastructures and practices in sharing scientific metadata
- A **poster session**, featuring a wide range of emerging work and interdisciplinary projects
- The presentation of the prestigious **Derek de Solla Price Award** by the international journal *Scientometrics*
- **Student Travel Awards**, supporting young researchers from around the world
- An award ceremony and closing events, celebrating contributions from across the global scientometric community, including:
    - The **Eugene Garfield Doctoral Dissertation Scholarship**, awarded to an exceptional doctoral student for outstanding research in the field
    - **Best Paper Award**, recognizing the most impactful and innovative research presented at the conference
- The **ISSI General Assembly**, where future plans and institutional developments were discussed

This volume publishes the peer-reviewed papers (*full papers, research-in-progress, and poster papers*) presented during the conference. It reflects the thematic richness and methodological diversity of our community, and highlights the increasing interrelation between scientometric methods and broader societal challenges — including sustainability, policy innovation, and the responsible use of AI.

We would like to express our deepest thanks to the International Society for Scientometrics and Informetrics (ISSI) for their trust and support, as well as to our academic partners, sponsors, and institutional collaborators in Armenia. Special appreciation goes to the reviewers, session chairs, keynote speakers, and the tireless members of the organizing and scientific committees.

Most importantly, we thank all authors and participants for their contributions to this vibrant intellectual exchange. May these proceedings serve as a valuable resource for ongoing research, and as inspiration for the continued development of scientometrics as a field committed to rigor, openness, and global inclusivity.

Shushanik Sargsyan, Wolfgang Glänzel, Giovanni Abramo

Index of proceedings papers

Full papers

Papers on the Main Paths are Associated with Lower Disruption in Scientometrics…….. 1301
Zizuo Cheng, Feicheng Ma

Quantifying the Political Attributes of Technology for Potential Bottleneck Technologies Identification: Evidence from Chinese Integrated Circuits Industry…… 1323
Tao Zhiyu, Liang Shuang, Li Hanxi, Liu Yajing

Quantitative Analysis of IITs' Research Growth and Contributions towards Achieving SDGs ……...1337
Parul Khurana, Kiran Sharma, Akshat Nagori, Manya, Mehul Dubey

R&D Innovation Patterns and Patent Application Strategy of Top-Selling Drugs: Insights from Patentometric ……... 1361
Chao-Chih Hsueh, Dar-Zen Chen

Rapid Growth of Research Output Amidst Political Instability: A Study of Libya's Last 2 Years ……1375
Stephen Wu, Adel Diyaf, Reem Abusanina

Research and Application on Multiple Topic Association Fusion Method Based on Neural Network and Evidence Theory ……...1389
Shuying Li, Xian Zhang, Jiahui Li, Xin Zhang, Haiyun Xu

Research on Public Perception of Academic Achievements under Public Health Emergencies …… 1415
Liu Xiaojuan, Hu Wei, Li Xinran, Xiao Yuntong

Research on the Innovation Performance Improvement Path of Scientific Research Team from the Perspective of Network Embeddedness……..1440
Junwan Liu, Qiqi Zhang, Shuo Xu, Chenchen Huang, Xiaoyun Gong, Jiahao Li

Research-Policy Alignment in AI: A Bibliometric Study of the EU AI Act…..1471
Cristian Mejia

Retracted Citations and Self-citations in Retracted Publications:
A Comparative Study of Plagiarism and Fake Peer Review ……. 1482
Kiran Sharma , Parul Khurana

Revisiting the field normalization approaches/practices…… 1493
Xinyue Lu, Li Li, Zhesi Shen

Role of Artificial Intelligence in Scientific Research:
Classification Framework and Empirical Insights ……. 1507
Zhe Cao, Yuanyuan Shang, Lin Zhang, Ying Huang, Gunnar Sivertsen

Science and Artificial Intelligence: A Copyright Perspective…… 1523
Dmitry Kochetkov

Science-Policy Tendencies in Armenia Towards
the International Collaboration …….. 1535
Gevorg Kesoyan, Aram Mirzoyan, Simon Hunanyan, Miranush Kesoyan

Scientific Travelers Associated with Less Disruption but
Better Scientific Novelty…… 1555
Mingze ZHANG, Penghui LYU, Yizhan LI, Zexia LI

Self Citations in Academic Excellence: Analysis of the Top 1% Highly
Cited India-Affiliated Research Papers ……. 1582
Kiran Sharma, Parul Khurana

Shaping Innovation: A Regional Perspective on Industrial PhD
Programs in Italy …… 1593
Tindaro Cicero, Annalisa Di Benedetto

Social Impact Analysis of Retracted Paper in the Context of
Public Health Emergencies …… 1613
Liu Xiaojuan, Shen Jianing, Dai Xinran, Yu Yao

Structures of Authors' Collaboration at Young Universities……. 1636
Nataliya Matveeva, Vladimir Batagelj

Study on the Differences Between Journal Papers and Conference Papers in
the Frontier of Basic Research: Taking the Terahertz Field as an Example….. 1660
Liu Hao, Chen Yunwei, Zhang Biao

Synergy Between Science And Technology In University-Industry Innovation Ecosystems: A Cross-National Comparison Of Elite Academic Partnerships In China, Germany, And The United States……. 1674
Hui Zhang, Hui Fu, Ying Huang

Text-based Classification of All Social Sciences and Humanities Publications Indexed in the Flemish VABB Database …….. 1699
Cristina Arhiliuc, Raf Guns, Tim C. E. Engels

The Effects of Research Evaluation: Do Researchers' Perceptions Align with Evidence? …… 1717
Giovanni Abramo, Ciriaco Andrea D'Angelo, Emanuela Reale, Antonio Zinilli

The Impact of Russia-Ukraine Conflict on International Migration of Russian-Affiliated Researchers ……. 1736
Andrey Lovakov

The Increasing Fragmentation of Global Science Limits the Diffusion of Ideas……1745
Alexander J. Gates, Indraneel Mane, Jianjian Gao

The Interaction between Scientific Research and Policy in The Field of Supply Chain: An Empirical Analysis Based on Overton Data ……1770
Li Jiangbo, Li Jiake, Mu Yingyu, Li Jian

The Trends of Open Access Academic Books and Discipline Dynamics: A Cross-database Comparison Based on OpenAlex and Web of Science …….1794
Li Jiangbo, Niu Shihang, Ouyang Wenhao, Li Jian, Zhang Mingyue

Trajectory of Research Method Usage in the Academic Careers of Scholars in the Library and Information Science ……. 1813
Jiayi Hao, Chengzhi Zhang

Transforming Researcher Evaluation: A New Global Platform to Measure Impact Across Disciplines ……..1841
Balázs Győrffy, Boglárka Weltz, István Szabó

Unveiling the Temporal Dynamics: The Impact of Knowledge Source Diversity, Breadth and Depth on Disruptive Innovation through Time-Series Analysis ……. 1849
Yue Li, Lele Kang, Jiaxing Li

Web Mining the Online Presence of Global Scientific Academies ……...1870
Xiaoli Chen, Xuezhao Wang

What Type of Methodological Novelty is More Disruptive? Evidence from Citation Classics 1899
Linlei Xie, Yi Zhao, Chengzhi Zhang

Where Did Post-Doctorates Go? A Factorized Analysis on Chinese Postdoctoral Program for Innovative Talent 1925
Tan Fu, Wen Lou

Research in Progress

A Dashboard to Visualize Retraction Statistics 1945
Ayush Tripathi, Achal Agrawal, Moumita Koley

A Hybrid Bibliometric-SEM-ANN Approach on Mapping the Intellectual Structure of Knowledge, Dynamic Capabilities, And Competitive Advantage 1952
Kuei Kuei Lai, Yu-Chun Hsu, Chwen-Li Chang

A Novel Type Collaboration: Global Big Science Facilities Co-utilization..... 1964
Zexia LI, Mingze Zhang, Lili Wang, Yizhan LI

Algorithmically Calculated Mentorship: The Netherlands Validation Study....1972
Kathryn O. Weber-Boer, Carlos Areia, Tamarinde Haven

Application of Molecular Docking Technology in Drug Discovery Based on Bibliometric and Patent Analysis 1978
Zhou Haichen, Jorge Gulín-González, Chen Yunwei

Automatic Literature Review Generation by Integrating Large and Small Models 1987
Xiaofei Li, Guo Chen

Beyond Sentiment Analysis with ChatGPT: Classifying Authors' Perspectives on Russian Topics 1995
Carolina Coimbra Vieira, Elena Chechik, Victoria Di Césare

Beyond Citations: Tracing and Validating the Rapid Adoption of AlphaFold in Biomedical Research Through Full-Text Analysis..... 2002
Haochuan Cui, Yuzhuo Wang, Kai Li

Biblium: An Advanced Python Library for Bibliometric and Scientometric Analysis ……2010
Lan Umek, Dejan Ravšelj

Book Authors as Self-Promoters on X (Twitter) and Their Information Dissemination Networks ……2035
Yajie Wang, Haiyan Hou, Alesia Zuccala

Catalytic Effect of Open Data Platforms: The Case of the Global Biodiversity Information Facility (GBIF) ……..2043
Honami Numajiri, Michio Oguro, Takayuki Hayashi

Conceptualizing Metascience Observatories ……. 2050
Emanuel Kulczycki, Johann Mouton, Didier Torny, Sergio Luiz Monteiro Salles Filho, Pei-Ying Chen, Juan Rogers, Noor Jaleel, Mathieu Ouimet, Kieron Flanagan, Aditi Ashok, Cassidy R. Sugimoto

Country Self-Preference and National Research Systems: A Path to Independence or Isolation? …… 2058
Jianjian Gao, Alexander J. Gates

Current Interdisciplinarity Measures Fail to Reflect Authors' Perspectives…... 2066
Dag W. Aksnes, Henrik Karlstrøm, Fredrik N. Piro

Difference between Preprint and Journal Systems ……2073
Chiaki Miura, Ichiro Sakata

Disciplinary Identity in the Origins of the Science of Science……... 2079
Emanuel Kulczycki, Przemysław Korytkowski

Distinguishing Types of Scientific Innovation Capacity: Exploring the Patterns and Dynamics of Knowledge Combinations and Impacts on Innovation in Biomedical Literature ……2087
Jinyu Gao, Yi Bu, Sarah Bratt

Distribution Differences of Knowledge Diversity among Authors in Different Contributor Roles—Evidence from 101014 PLOS ONE Articles…..2095
Jingyuan Li, Yi Zhao, Jiaqi Zeng, Chengzi Zhang

Does Distance Still Matter? The Impact of Geographic Patterns in Scientific Knowledge Sourcing on Invention Value ……..2103
Guiyan Ou, Chaocheng He, and Jiang Wu

Enhancing Scientometrics Prediction under Uncertainty:
A DIKW-Based Framework and Methodological Synthesis ……2112
Shuya Chen, Guo Chen

Exploring Google Books, Open Library, and Wikipedia as
Sources for Book Metadata: The UK and Lithuanian Cases ……2121
Eleonora Dagienė

Exploring Research Collaboration of Private Universities in Emerging
EU Countries: A Comparison with Public Sector ……..2130
Alexander Dmitrienko, Nataliya Matveeva, Maria Yudkevich

Exploring the Policy Impact and Funding Mechanisms of Scientific
Collaboration Between Taiwan and New Southbound Policy (NSP)
Priority Countries ……...2138
Pei-Ying Chen, Tzu-Kun Hisao, Cassidy R. Sugimoto

Field Differences in External Funding: An Analysis of Funding
Composition of Externally Funded Publications ……... 2146
Fredrik Niclas Piro, Henrik Karlstrøm, Ida Svege, Dag W. Aksnes

Geographies of Underrecognition: Citation Disparities in Russian Studies…... 2154
Katerina Guba, Elena Chechik, Angelika O. Tsivinskaya, Artur Pecherskikh, Nikita Buravoy

Harnessing Data Papers: An Analysis of Their Role in Scientific Data
Dissemination and Reuse ……... 2161
Liyue Chen, Xiaomin Liu

Exploring the Effects of Migration for Social and Humanity Researchers…... 2169
Alena Nefedova, Elizaveta Chefanova

How is the Sino-US AI Collaboration Reshaped by the China Initiative?..... 2177
Dingkang Lin, Jiang Li

How systematic are the systematic reviews? ……2185
Andrey Guskov, Denis Kosyakov, Irina Selivanova, Alexandra Malysheva

Impact of Marriage on Productivity and Career of Women Scholars ……2193
Shiqi Tang, Xianjiang Deng, Jianhua Hou, Cassidy R. Sugimoto

Implicit Reporting Standards in Bibliometric Research: What Can Reviewers' Comments Tell Us About Reporting Completeness?..... 2199
Dimity Stephen, Alexander Schniedermann, Andrey Lovakov, Marion Schmidt, Matteo Ottaviani, Nikita Sorgatz, Roberto Cruz Romero, Torger Möller, Valeria Aman, Stephan Stahlschmidt

Is Journal Citation Indicator A Good Metric for Art & Humanities Journals Currently? …….. 2207
Yu Liao, Li Li, Zhesi Shen

Mapping and Quantifying the Boundaries in Research Data Sharing based on Data Policy ……2214
Yizhan LI, Mingze ZHANG, Lu DONG, Zexia LI

Mapping the Social Structure of Philosophy of Science Through Large-Scale Acknowledgments Analysis ……2228
Eugenio Petrovich, Edoardo Fazzini, Lorenzo Gandolfi

Melting Science: Russian Climate Change Research in the Global Context ….2237
Alexey Zheleznov, Ekaterina Dyachenko, Maxim Dmitriev, Katerina Guba

Missing links in the chaîne opératoire of citation: the limitations of systematic literature search in the social sciences and humanities……2244
Kathryn O. Weber-Boer

National Mobility and Career Performance of the Scientific Workforce in Colombia ……..2251
Jesús María Godoy, Yajie Wang, Julián D. Cortés

National research, national policy: how local research fuels Brazil's policy… 2259
Bernardo Cabral, Evandro Cristofoletti, Karen Esteves Fernandes Pinto, Sergio Salles-Filho, Yohanna Juk

Not All 'Predators' are the Same: Exploring the Spectrum of Questionable Journals ……..2267
Zehra Taşkın, Güleda Doğan, İdris Semih Kaya, Ezgi Uğurlu, Özge Söylemez, Ceren Bilge Seferoğlu, Emanuel Kulczycki

Old but Not Obsolete: Bag-of-Words vs. Embeddings in Topic Modeling …...2275
Jean-Charles Lamirel, Francis Lareau, Christophe Malaterre

Originality in Scientific Titles and Abstracts Can Predict Citation Count …… 2283
Jack H. Culbert, Yoed N. Kenett, Philipp Mayr

Proposal of INDIRECT X Mentions as an Altmetrics Indicator:
Dissemination of Research Papers on X via Web News and Blogs…….. 2291
Ai Kishimoto, Takayuki Hayashi

Regional Patterns of Plagiarism: Evidence from PhD Theses in Russia……. 2299
Anna Abalkina, Alexander Libman, Andrey Zayakin

Research leadership recommendation in research leading-participating
multiplex networks based on Wasserstein Distance …….. 2306
Chaocheng He, Guiyan Ou, Fuzhen Liu, Sitong Xiang, Ye Zhang, Jiang Wu

Research on the Measurement Method of Disciplinary Diversity
Based on Lexical Semantic Analysis ……. 2325
Guo Chen, Yifan Yang

Single Authorship, National Co-Authorship, and International
Co-Authorship in the Social Sciences and Humanities:
A Multi-Dimensional Analysis of the Flemish Case ……. 2333
Peter Aspeslagh, Tim C.E. Engels

Small Open Access Publishers: An Analysis of Visibility and Impact Patterns. 2340
Roberto Cruz Romero

Structural and Institutional Determinants of Open Access Publishing:
A Macro-Perspective ……..2349
Roberto Cruz Romero, Stephan Stahlschmidt

Unraveling the Driving Factors of Team Performance: The Impact of Team
Composition and Collaboration Relationships on Project Teams…….. 2359
Ruinan Li, Tingcan Ma, Beibei Sun, Yuzhuo Wang

Unveiling Tortured Phrases in Humanities and Social Sciences……... 2367
Alexandre Clausse, Fidan Badalova, Guillaume Cabanac, Philipp Mayr

What are the Most Important Elements of Research Activity to Assess?
The Proposal of Relational Goods ……. 2374
Cinzia Daraio, Antonio Malo, Giulio Maspero, Ilaria Vigorelli

Will Scientific Research Drive Technology to be a Hit? A Comparison between
Emerging Technological Fields and Traditional Technological Fields …… 2387
Xi Chen, Jin Mao, Gang Li, Xuehua Wu

Posters

15 Years of the Eastern Partnership initiative: a Bibliometric Reflection……. 2397
Maria Ohanyan, Mariam Yeghikyan, Miranush Kesoyan, Simon Hunanyan, Aram Mirzoyan

A Framework for Analyzing Identification Funds in the Social Sciences under the Perspective of Country Mentions: An Example of China and the United States ….. 2400
Changcheng Xue, Kaiwen Shi, Hongyu Wang, Xiaoguang Wang

A look behind metrics for knowledge integration: Some notable cases …… 2403
Pei-Shan Chi, Wolfgang Glänzel

AI-Powered Evaluation of Peer Review Quality: A Case Study of eLIBRARY.RU …….. 2406
Dmitry Kochetkov, Denis Kosyakov, Irina Lakizo, Viktor Glukhov, Andrey Guskov

Analysis of compliance with the FAIR principles in Education Science…….. 2409
Andrea Sixto-Costoya, Adolfo Alonso-Arroyo, Luiza Petrosyan, Rafael Aleixandre-Benavent, Rut Lucas-Domínguez

Attempts to Enable Generative AI for Topic Recognition: A Case Study of ChatGPT ……. 2412
Wenting Tang, Wen Lou

Automating Reproducible Bibliometrics with the Open Research Converter... 2415
Jack H. Culbert, Philipp Mayr

Can Large Language Models Accurately Discriminate Subject Term Hierarchical Relationship? ……... 2418
Yuanxun Li, Hongyu Wang, Kaiwen Shi, Xiaoguang Wang

Delayed Recognition of Novel Ideas: Initially Underestimated, Ultimately Rewarded ……. 2421
Tao Zhiyu, Liu Xiaoping, Liang Shuang, Li Hanxi

Does winning an Ig-Nobel Prize have an impact on the visibility of the winners' research work? …….. 2424
Philippe GORRY

Enlarging the spectrum. Implementing a local extension of ROR as identification instrument for additional actors in Flemish (SSH) research….. 2427
Peter Aspeslagh

EU-Armenia Scientific Partnership: A Bibliometric Analysis of Funding and Academic Output ……2430
Ruzanna Shushanyan, Maria Ohanyan, Miranush Kesoyan, Mariam Yeghikyan, Gevorg Kesoyan

Evaluating Large Language Models for Gender Bias in Academic Knowledge Production …… 2433
Judit Hermán, Kíra Diána Kovács, Yajie Wang, Orsolya Vásárhelyi

Exploring institutional type composition in scientific collaboration and its role in scientific impact ……2436
Shuang Liang, Zhiyu Tao, Qingshan Zhou

Fully Algorithmic Librarian: Large-Scale Citation Experiments …….. 2439
Tomasz Stompor, Janina Zittel, Thorsten Koch, Beate Rusch

Fundamental Foundations of Scientific Heritage Formation: The Evolution of Scientific Knowledge and the Possibility of Applying Scientometric Tools ….2442
Victor A. Blaginin, Elizaveta V. Sokolova

Fusing Multi-Source Data through a Multi-Layer Network for Technological Opportunity Identification ……..2445
Jinzhu Zhang, Mingxia Lu, Haoyu Li

Gender Disparities in Editorial Board Member in Information Science & Library Science Journals …….. 2448
Yiming Liu, Rut Lucas-Domínguez, Adolfo Alonso-Arroyo, Cristina Rius, Rafael Aleixandre-Benavent

Gender Leadership in Cancer Research ……2451
Cristina Rius, Yiming Liu, Adolfo Alonso-Arroyo, Rafael Aleixandre-Benavent, Rut Lucas-Domínguez

Healthcare Cybersecurity: Insights from a Scientometric Approach……. 2454
Simone Di Leo, Cinzia Daraio, Fabio Nonino, Eugenio Oropallo

How Does Knowledge Source Novelty Influence Knowledge Output Novelty? Evidence from 269,569 PLOS Articles ……..2457
Yi Xiang, Chengzhi Zhang

How far are we from understanding phygital healthcare convergence? Building an AI knowledge map grounded in bibliometric metadata……. 2460
Cinzia Daraio, Simone Di Leo

India and quantum computing-related publications …….. 2463
Xiaojun Hu, Ronald Rousseau

Machine learning-based model to predict topics contributing to Sustainable Development Goals: A study of Latin American and European Countries…… 2466
Barbara S. Lancho Barrantes

Measuring the effect of research award on collaboration relationships……..2469
Chien Hsiang Liao

MetaInfoSci: Visualize trends and understand facts ……..2472
Kiran Sharma, Parul Khurana, Ziya Uddin

Multidimensional quantitative analysis of the fit of Chinese science and technology talent policy ……2475
Wang Kaile, Chen Yunwei

Multilevel Structures, Connection and Balance:
The Evolution of the Structure of Science ……2478
Yuxian Liu, Hongrui Yang, Ronald Rousseau, Raf Guns, Sisi Li, Yafang Fan, Helan Wu, Sanfa Cai

Open Citations in German Educational Research–Identifying Disciplinary Practices to Train Data Extraction ……2481
Verena Weimer, Tamara Heck, Christoph Schindler

Portuguese Scientific Production: Volume indicators ……... 2484
Catarina Carreira, Cristiana Agapito

Productivity and Impact Patterns in Scientific Careers …….. 2487
Kaile Gong

Publications at the Intersection of Academia and Market:
Unpacking Scholarly Outputs of University-Industry Collaboration in Brazil.. 2490
Gabriel Falcini, Sergio Luiz Monteiro Salles Filho, Yohanna Juk

Quality Evaluation of Scientific Journals in the Open Science Context ……2493
Lei Li, Yue Hu, Hui Peng, Shi Chen

Recent Advance of Text Mining in LIS: A bibliometric review ……... 2496
Siqi Hong, Guo Chen

Research Collaboration and Leading Role:
A Comparative Study on the Academic Communities in Japan and Taiwan…. 2499
Szu-chia Lo, Yuan Sun

Research on Scientific Frontier Topics Based on Citation Analysis and Content Analysis - Taking the Structural Analysis of Nature Index as an Example……2502
Tan Xiao, Li Hui, Xu Haiyun, Li Jiayu, Jin Xiaohong, Xi Guiquan, Zhang Ting, Chen Shu

Research on the Path of China's Construction of a World Science and Technology Power …… 2505
Wang Kaile, Chen Yunwei

Revolutionizing Medical Processes Through Phygital Technology: A Multiple Case Study Approach ……..2508
Eugenio Oropallo, Cinzia Daraio, Simone Di Leo, Fabio Nonino

Single journal bibliometric case studies ……2511
Ilya Gorelskiy, Daniel Karabekyan, Alexander Karpov

Stop, little pot! Are there too many scientometric studies? ……...2514
Ekaterina Dyachenko, Alexey Zheleznov, Maxim Dmitriev, Katerina Guba

The Effectiveness of Large Language Models in Predicting User Question Preferences on ResearchGate Q&A ……...2517
Lei Li, Yue Hu, Hui Peng

The Impact of Brazilian Scientific Production on Public Policies: A Scientometric Analysis ……...2520
Bernardo Cabral, Carlos Graziani, Evandro Cristofoletti, Guilherme Macari, Karen Esteves Fernandes Pinto, Sergio Salles-Filho, Yohanna Juk

Trends and Distribution of Domestic and International Research Collaboration: An Asian View …….. 2523
Szu-Chia S. Lo, Mu-Hsuan Huang

Unraveling Evolutionary Dynamics of Disruptive Innovations: Insights from Multi-Scale Knowledge Networks…….. 2526
Haiyun Xu, Junhao Yang, Xiao Tan, Shuying Li, Zenghui Yue, GuotingYuan, Xin Li

Author Index…….. 2529

FULL PAPER

https://doi.org/10.51408/issi2025_009

Papers on the Main Paths are Associated with Lower Disruption in Scientometrics

Zizuo Cheng[1], Feicheng Ma[2]

[1] *chengzizuo@whu.edu.cn*, [2] *fchma@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan (China)

Abstract

We integrate two bibliometric frameworks—the Disruption Index (DI) and Main Path Analysis (MPA)—to examine how scientific papers shape knowledge flows in scientometrics. The DI measures a paper's capacity to shift citation patterns: a positive DI indicates the paper "diminishes" its predecessors (disruptive impact), while a negative DI suggests it reinforces prior work (consolidative impact). The MPA identifies dominant knowledge trajectories by extracting the most frequently traversed citation paths within a field, highlighting papers critical for sustained knowledge transmission. Analyzing 36,523 scientometrics publications, we find papers on main paths exhibit lower disruption, with disruption declining further over time. It aligns with MPA's tendency to amplify consensus-driven knowledge. Disruptive papers (DI>0) are less likely to appear on main paths, suggesting alternative diffusion pathways. Besides, indirect impact metric (SPX) is positively associated with direct impact (citation counts) but negatively correlated with disruption. Our research shows that MPA may underrepresent disruptive contributions, necessitating complementary DI/SPX evaluation.

Introduction

Information scientists aim to use citation relationships to identify impactful scientific papers. Citation count is the most common evaluation metric for its simplicity and intuitiveness. However, it overlooks the complex information within citation structures (Bu, Waltman, & Huang, 2021). Recently, the disruption index (DI) proposed by Funk and Owen-Smith (2017) has garnered significant attention (Wu, Wang, & Evans, 2019; Park, Leahey, & Funk, 2023; Lin, Frey, & Wu, 2023; H. Li, Tessone, & Zeng, 2024). Unlike citation count, DI focuses on measuring the nature of a paper's impact (Leahey, Lee, & Funk, 2023). It assesses a paper's influence based on how it disrupts existing citation patterns: when subsequent papers cite a focal paper (FP) but do not acknowledge FP's references, FP disrupts its field; conversely, FP consolidates the field's development (Azoulay, 2019). In other words, the FP's brilliance captures the attention of successors and dims its predecessors. Scholars have examined DI's validity through expert evaluations (Bornmann & Tekles, 2019; Bornmann et al., 2020a, 2020b). Some researchers explored Nobel Prize-winning papers, which often have both high DI values and citation counts(Liang, Lou, & Hou, 2022). These two metrics, reflecting the nature and level of impact, provide a two-dimensional evaluation framework (Wei, Li, & Shi, 2023). In this framework, most papers contrast with Nobel Prize-winning works, exhibiting lower citation counts and DI values. The remaining papers fall into two categories. A high DI value does not equate to a significant impact, as these papers might receive fewer citations. Conversely, highly cited papers may not possess high DI values. Review articles exemplify this, as they primarily integrate existing knowledge.

These combinations capture our interest, the two with high impact levels. Citation relationships represent a form of knowledge flow, and the DI measures how FPs disrupt this flow. Papers with high citation counts often play crucial roles in knowledge flow and tend to cluster along the main paths of citation networks. Scholars have analyzed these main paths to map the development of fields (Hummon & Dereian, 1989) and identify foundational papers (Ma & Liu, 2016). However, these studies often overlook whether the knowledge flow reflects disruption or integration. Introducing DI can help us analyze how papers on the main paths contribute to knowledge flow within specific fields.

We have additional motivation for using main path analysis (MPA). While citation count reflects the direct impact of a paper, it fails to capture indirect influence. MPA offers a complementary measure (Liu, Lu, & Ho, 2019). Furthermore, both the DI and MPA consider FP's citing and cited papers, aligning them conceptually. The DI focuses on local network structures, whereas MPA utilizes global information. Integrating network information may better measure a paper's impact, allowing us to develop a three-dimensional evaluation framework. Given the rapid growth in scientific publications, using larger datasets to represent specific research fields is essential but challenging. MPA can guide us in focusing on a subset of papers that can effectively represent the core of the research field.

We select scientometrics as a case study to address the following research questions. First, do papers on the main paths exhibit higher disruption? Do the disruptive papers tend to appear on the main paths? Second, is the indirect impact measure associated with MPA consistent with other paper evaluation metrics?

Literature Review

Disruption Index (DI)

Researchers have conducted in-depth discussions on the DI. Here we only provide a brief overview of this index. We can refer to Leibel and Bornmann (2024) for a more detailed one. To facilitate subsequent elaboration, we first introduce the regular form of this index.

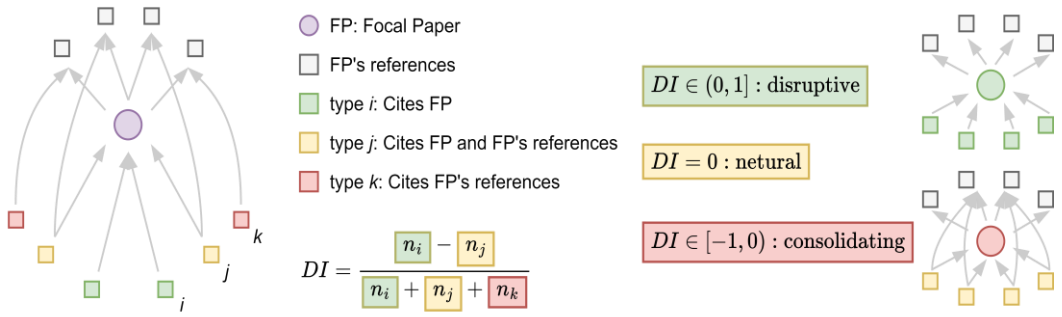


Figure 1. Illustration for DI.

In Figure 1, for an FP, we focus on its references and citing papers. It has four references (gray rectangles). The six citing papers (rectangles below) are in three

parts: those citing only the FP (green, denoted as i), those citing both the FP and its references (yellow, denoted as j), and those citing only its references (red, denoted as k). The DI value for the FP is the difference in proportion between the i and j, i.e.,

$$DI = p_i - p_j = \frac{n_i - n_j}{n_i + n_j + n_k} \in [-1,1]$$

We consider $DI = 0$ as a threshold. $DI > 0$ indicates the paper is disruptive, while $DI < 0$ suggests it is consolidating. Additionally, we should determine the number of citing papers, which requires setting an appropriate citation window.

The first type of research examines the DI mechanism. Leydesdorff and Bornmann (2021) argue that the DI relies on bibliographic coupling, where the coupling of the FP and its references signifies continuity, while disruption indicates a break in continuity. Lin, Evans, and Wu (2022) suggest that disruptive papers often achieve breakthroughs in theory, methods, or discoveries compared to their references.

Further discussions on improvements are in two factions. One faction views the DI as a relative measure, considering disruption and integration as opposing concepts. The other treats it as an absolute measure, calculating disruption and integration separately (Chen, Shao, & Fan, 2021; Leydesdorff, Tekles, & Bornmann, 2021). Current research focuses more on the former approach. Since many k-type papers can skew the DI value towards zero, Bornmann et al. (2020b) propose setting a bibliographic coupling threshold to reduce the number of k-type papers. Deng and Zeng (2023) suggest severing links between citing papers and highly cited references to increase the number of i-type papers. Both methods adjust the DI value. Ruan et al. (2021) note that fewer references negatively impact the DI value and recommend focusing only on FPs with more than ten references. Yang et al. (2024) systematically review the shortcomings of the DI and offer more reasonable modifications. Yang et al. (2024) also propose disruptive citations to measure a paper's absolute disruptive impact.

Validation work relies on specific datasets, including milestone paper lists published by *Physical Review Letters* in physics (Bornmann & Tekles, 2021) and peer review results from F1000Prime in biology and medicine (Bornmann et al., 2020b). A notable validation effort is Macher, Rutzer, and Weder's rebuttal (2024) of Park, Leahey, and Funk's conclusions (2023), highlighting that truncating the citation window can lead to biased results.

The second type of research examines how research activities impact the papers' disruption. Lyu et al. (2021) show that team size and international collaboration negatively correlate with the papers' DI value. Zeng et al. (2021) report a positive correlation between new teams and the papers' disruption. Wang et al. (2023) reveal that scientists in structural holes within collaboration networks are more likely to publish disruptive papers. Zhao et al. (2024) note that teams with more thought leaders produce less disruptive ideas. Another set of studies investigates the impact of interdisciplinary collaboration on paper disruption. Liu et al.'s empirical results (2024) indicate that collaboration within the same discipline is more likely to produce disruptive outcomes, while Chen et al. (2024) present research with opposite conclusions. Other influencing factors include funding types (Yang & Kim, 2023) and prior knowledge (Sheng et al., 2023).

The third type of research expands the DI application scenarios. Scholars use it for scientific evaluation, applying it to papers (Zhou et al., 2022; Wang et al., 2024; Yan & Fan, 2024a), scientists (Wang, Zhou & Zeng, 2023; Yang et al., 2023), and journals (Jiang & Liu, 2023).
Overall, researchers primarily focus on the first type of research. Future research may explore using textual information to measure paper's disruption and enhance the utilization of this index.

Main Path Analysis (MPA)

MPA is a classical network method that considers citation relationships as knowledge flows, tracing the most significant dissemination paths within a field. It involves two steps: calculating the traversal weights of links and extracting the paths with the highest weights. Current research focuses on methodological improvements to achieve more interpretable results.
Early explorations focus on network topology. Hummon and Dereian (1989) establish the foundation for MPA by proposing three traversal weight methods: Node Pair Projection Count (NPPC), Search Path Link Count (SPLC), and Search Path Node Pair (SPNP). Batagelj (2003) introduces the Search Path Count (SPC), which balances inflow and outflow traversal weights. Although SPC was initially popular, Liu, Lu, and Ho (2020) conclude that SPLC better suits the knowledge dissemination context after comparing the four methods. In path searching, Liu and Lu (2012) propose the main paths: local, global, and key-route. Additionally, Pajek (Everton et al., 2018) significantly contributes to disseminating MPA, offering researchers convenience. Researchers also explore other perspectives. For instance, Liu and Kuan (2016) examine the decay of knowledge during the flow process, Jiang, Zhu, and Chen (2020) address MPA's limitations in self-loop networks, Ho, Liu, and Chang (2017) investigate the impact of review papers on generating main paths, and Kuan analyzes MPA's tendency toward long path results (2023), proposing quantitative methods to evaluate main paths (Kuan & Liao, 2024).
Subsequent studies emphasize the integration of semantic information. For example, Chen et al. (2022) introduce link semantic weights to improve paths thematic coherence. Yan and Fan (2024b) incorporates knowledge graphs to enhance the knowledge proximity of path nodes. Additionally, Liu, Lu, and Ho (2019) suggest using link traversal weights to measure the indirect influence of papers within a field, although this idea has received limited attention.

Methods

Data Collection and Network Construction

Constructing a citation network includes two steps: determine the paper set in scientometrics and obtain citation relationships (their references and citing papers). We have two accessible data sources: Web of Science (WoS) and OpenAlex (Priem, Piwowar, & Orr, 2022).
For the first step, Bornmann and Tekles (2019b) select papers from *Scientometrics* to represent this field. Both WoS and OpenAlex offer a retrieval tool that uses the

Leiden algorithm (Traag, Waltman, & Van Eck, 2019) to cluster papers and assign category labels, which facilitates our research. Therefore, we obtain data separately and compare them. The strategy is in Table 1.

Table 1. Retrieval strategy for papers in scientometrics.

Source		*Strategy*
WoS	Query	TMSO= (6.238 Bibliometrics, Scientometrics & Research Integrity)
	Index	SCI & SSCI
	Document	Article & Review
	Date	2024-12-18
	Records	40,500
OpenAlex	Query	Topic is "scientometrics and bibliometrics research"
	Document	Article & Review
	Records	51,690

The results show that OpenAlex provides more data, and only 8,029 entries overlap, indicating significant differences. Merging the two datasets is feasible, but we are concerned that it could introduce more noise. Therefore, we manually check some classic papers in scientometrics. For instance, in "An index to quantify an individual's scientific research output," Hirsch proposed the famous h-index (2005). However, OpenAlex categorizes this paper under "Cognitive Science and Mapping." Clustering algorithm may bring noise especially when the data is large and complex. Considering the data quality, we prefer the WoS data. We also acknowledge the limitation of the manual review, conducting experiments separately may be a better choice.

For the second step, we choose the OpenAlex data. First, early papers often have limited references, and WoS does not index them. It may affect the DI value of papers. Additionally, WoS does not provide bulk access to forward citations, making it challenging to construct a complete network when the FP set is large. However, OpenAlex assigns universal identifiers (OpenAlex ID) and provides powerful APIs, overcoming the shortcomings abovementioned.

Overall, we finally adopt a mixed strategy: WoS provides focal paper set and OpenAlex offers citation relationships.

The comparison between WoS and OpenAlex data is in Table 2. For the 40,500 records, OpenAlex indexes most of them. Besides, OpenAlex covers 70% of the reference data for WoS and provides more references.

Table 2. Data Comparison between WoS and OpenAlex.

	WoS	*OpenAlex*	*Shared*
Records	40,500	40,182	40,182
Reference Items	267,803*	384,627	187,398
References	757,379	1,171,004	553,012

* Only 258,580 items exist in OpenAlex.

We have two citation networks in this study. First, we utilize complete data to construct a full network. Here, we select a portion of the 40,182 original records with at least one reference for the FP set. It covers 676,140 nodes and 6,568,462 edges. We also build a close network which only retains citations where both sides belong to the FP set (Li & Chen, 2022). The illustration is in Figure 2. Such an approach reflects knowledge flow within scientometrics, aligning with the strategy used in MPA studies. Table 3 shows the minor differences between the two close networks.

Table 3. Close network comparison between WoS and OpenAlex version.

	WoS	*OpenAlex*	*Shared*
Nodes	35,319	36,523	33,438
Edges	388,588	376,958	349,561

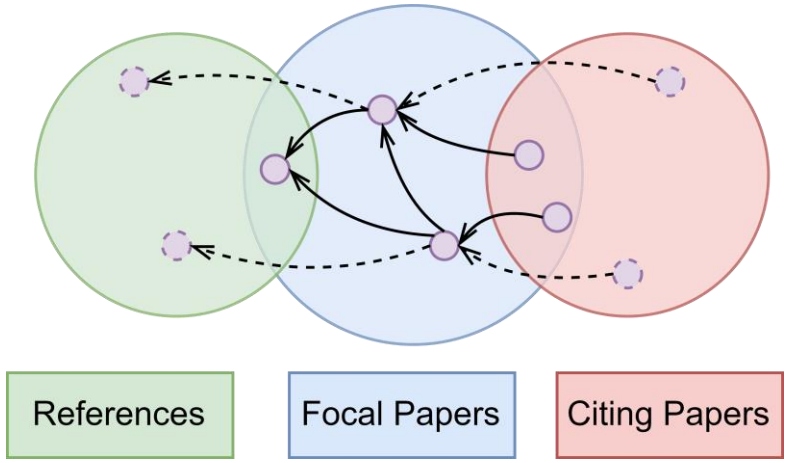


Figure 2. Illustration for close network construction.

Evaluation Metrics

The metrics we select to evaluate paper's impact are in Table 4.

Table 4. Evaluation metrics.

Dimension	*Metrics*	*Illustration*
Level	I_{10}	Citation counts within a 10-year citation window.
	I_{2024y}	Citation count received until 2024.
Nature	DI_{10}	DI within 10-year citation window.
	DI_{2024y}	DI values in 2024.

Main Path Analysis with Indirect Impact Metrics

We choose SPLC to calculate citation traversal count because this is more consistent with the representation of knowledge flow (Liu, Lu, & Ho 2019). We use multiple methods integrated in Pajek to extract the main paths for comprehensive results (Liu & Lu, 2012). We accomplish the task only on the close network to reduce bias (Filippin, 2021).

We also refer to the method provided by Liu, Lu, and Ho (2019) to measure the paper's indirect impact. Each FP has n citation links whose sum of citation traversal counts is s, and its indirect impact is $SPX = \frac{s}{n}$. The illustration is in Figure 3.

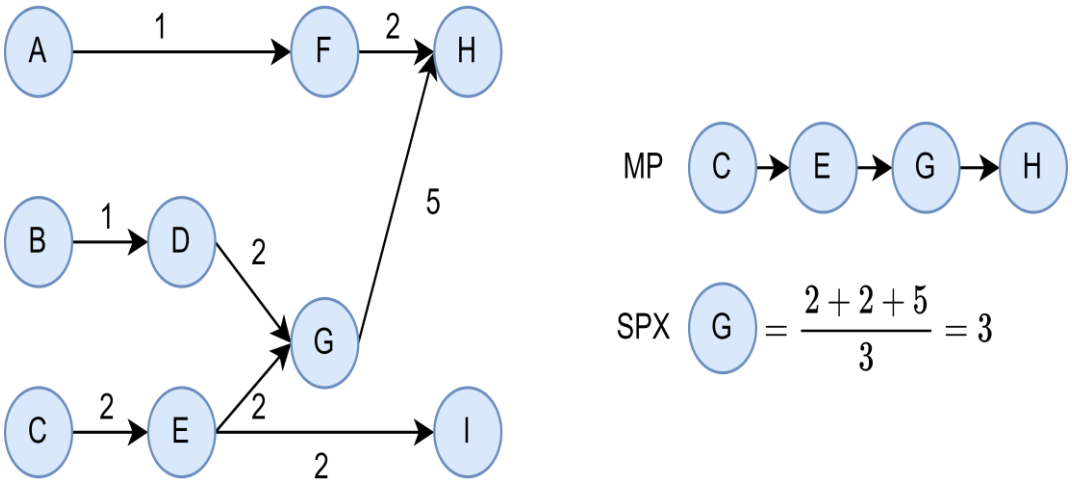


Figure 3. Illustration for MPA and SPX metrics.

Results

Network Description

In the citation network, the in-degree represents the citation count, and the out-degree represents the number of references. Figure 4 illustrates the logarithmic distribution of the FPs in the two networks.

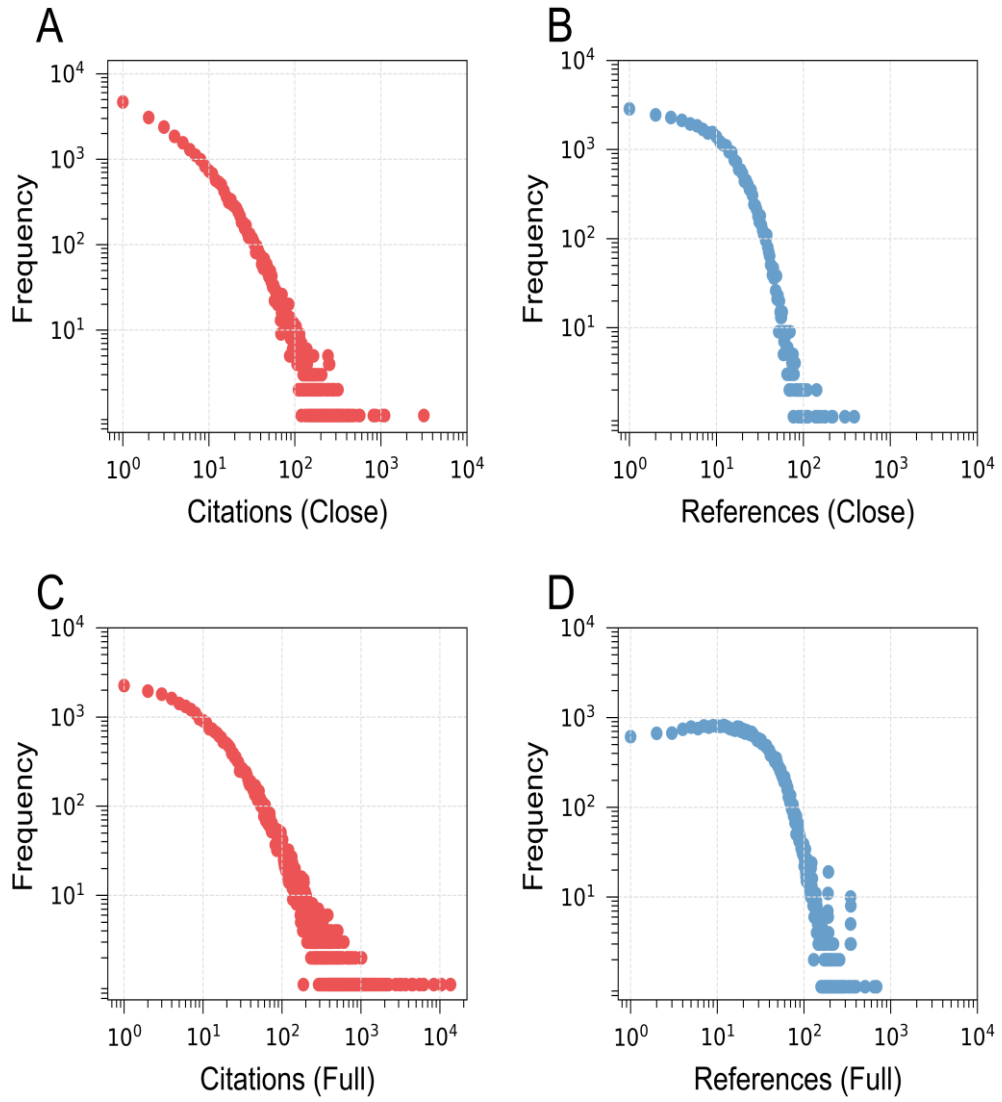


Figure 4. Citation and reference distribution in the two networks.

Most papers have less than 10 citations, while a few obtain extremely high impact. Hirsch's proposal on the h-index receives significant attention in the close network. In the full version, Van Eck and Waltman (2010) have the highest impact with the introduction of VOSviewer. One likely reason is that VOSviewer has become fundamental to scientometrics, leading researchers in the field to choose not to cite it. The distribution of reference is more concentrated in the upper range. Earlier papers tend to have fewer references, and OpenAlex may not fully index them. The paper with the most references is a 2008 review by Bar-Ilan (2008). In the full network, some nodes with numerous references, such as "Quantitative Studies of Science: A Current Bibliography" (ID "W2135332121"), appear. OpenAlex sometimes provides extensive but incorrect reference relationships for these nodes, introducing noise into the network.

Park et al. suggest that the current decline in the disruption of papers may be due to researchers bearing a heavier knowledge load (2023). Figure 5 presents a box plot showing the distribution of reference counts for FPs published from 1955 to 2024. Over time, researchers in scientometrics have consulted more literature.

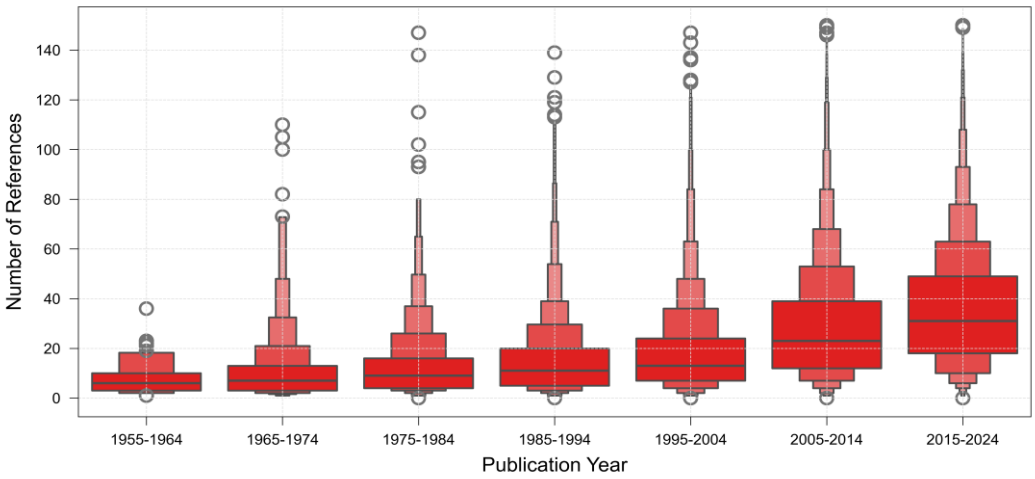


Figure 5. Reference distribution over years in the full network.

We examine the temporal distribution of citation behaviour. Figure 6 illustrates that most FPs derive insights from works published within the last decade and receive citations within ten years of publication. The citation window influences both the citation count and disruption. Thus, setting the ten-year window is more proper in this context.

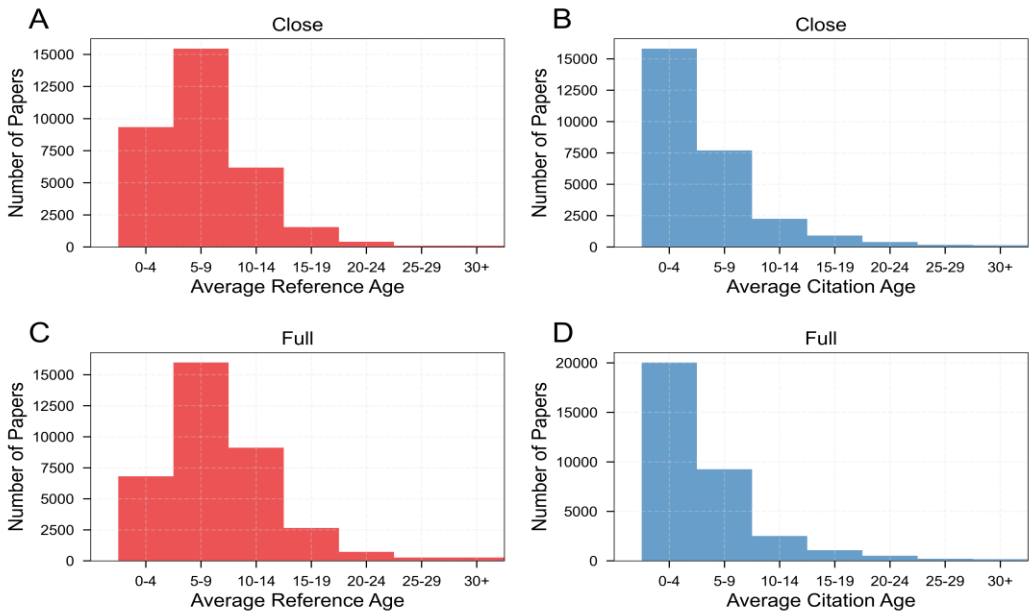


Figure 6. Average reference and citation age distribution in the two networks.

Disruption Distribution

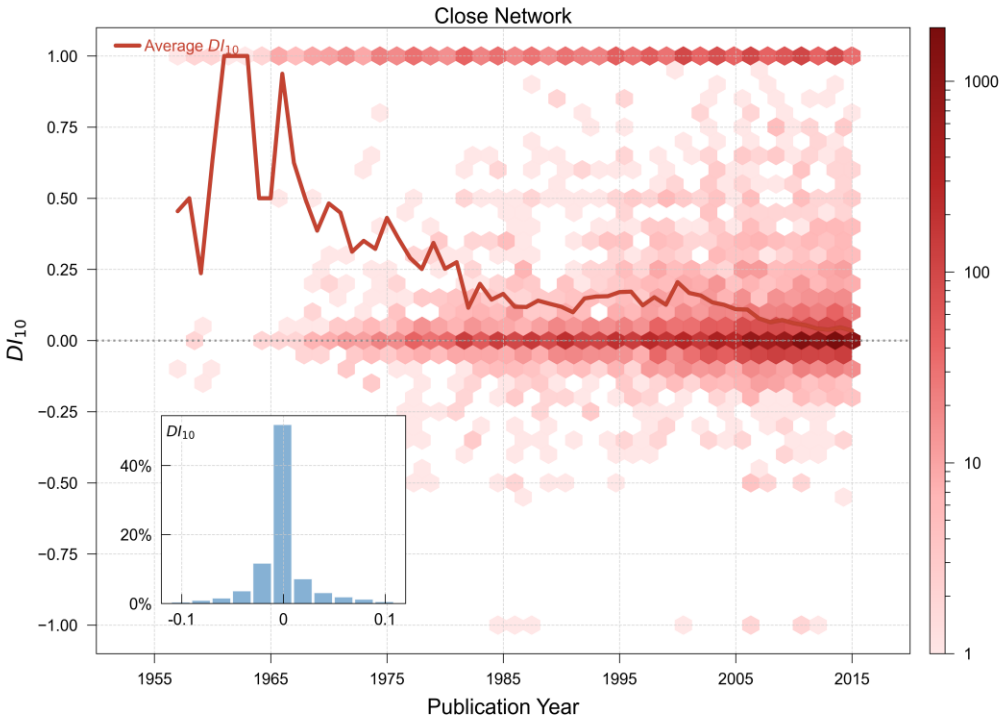


Figure 7. Distribution of DI_{10} values in the close network.

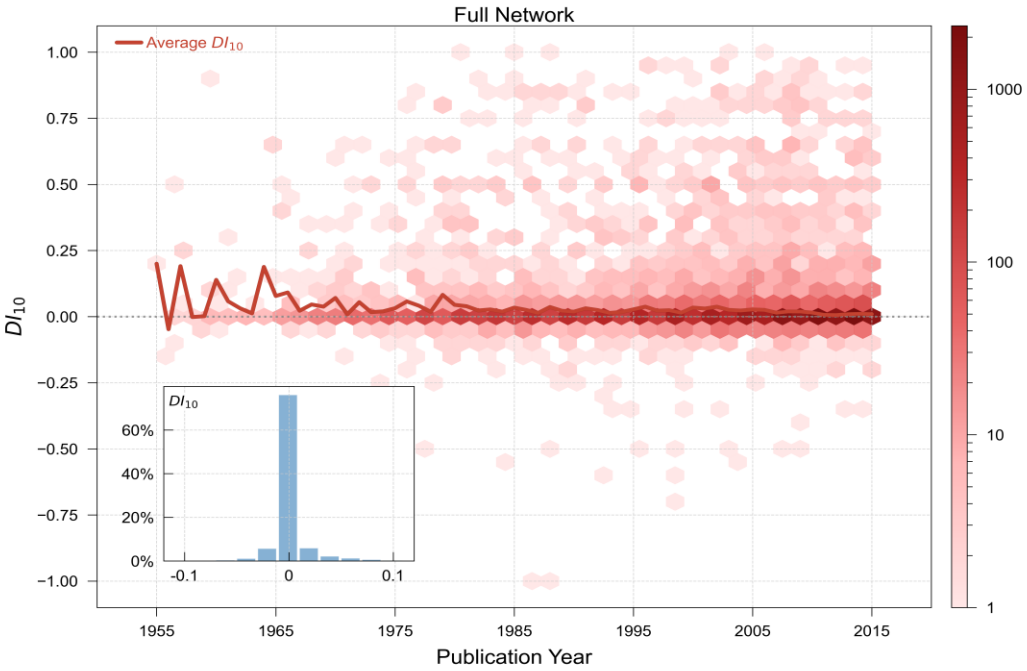


Figure 8. Distribution of DI_{10} values in the full network.

Figure 7 illustrates the distribution of DI_{10} for papers published in the close network from 1955 to 2014. A total of 1088 nodes are absent due to a denominator of zero. The main part is a hexbin plot, where each hexagonal area corresponds to a specific publication year and DI_{10} value, with colour indicating the density of papers. Most papers have DI_{10} values concentrated around zero. The histogram in the lower left corner reflects a similar trend, with over 40% of papers having a DI_{10} value of zero. The line graph depicts the average DI_{10} value per year, suggesting that the field of scientometrics is experiencing a decline in disruption. Figure 8 presents a similar picture, showing even lower annual average DI_{10} values and a more extreme distribution.

In the close network, we see variations in DI_{10} values. Figure 9 illustrates this dynamic. Red nodes have higher DI_{10} values in the full network, while blue nodes appear more disruptive in the close network. Blue nodes in the grey area show extremely high DI_{10} values, indicating that the structure of the close network significantly impacts these measurements.

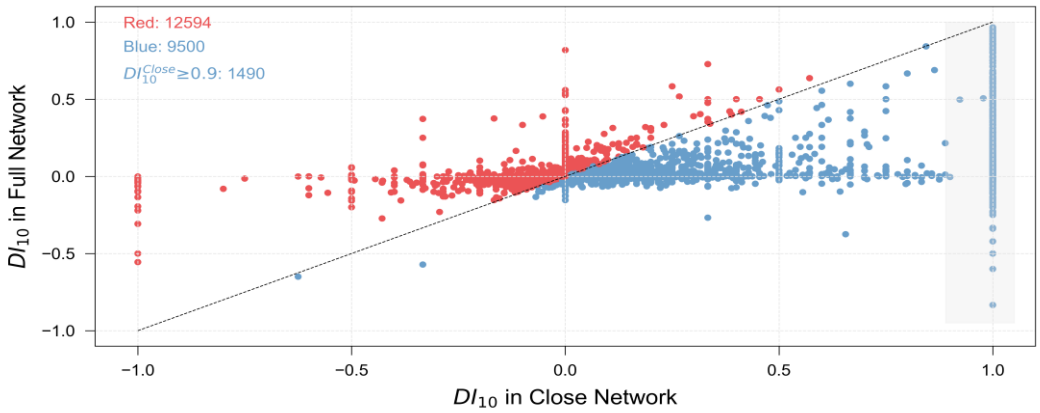


Figure 9. Distribution of DI_{10} difference between the two networks.

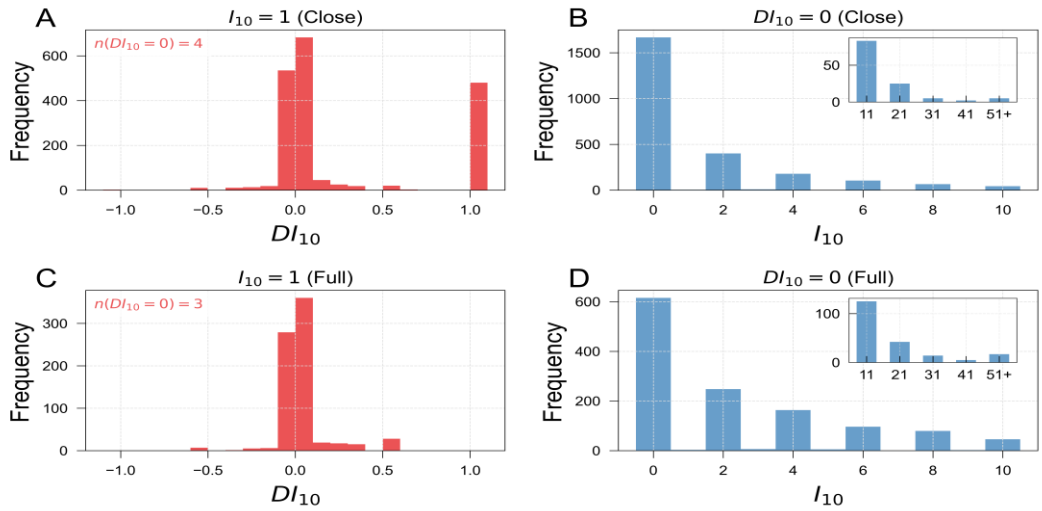


Figure 10. Distribution of DI_{10} and I_{10} values for papers with $I_{10} = 1$ or $DI_{10} = 0$.

We combine I_{10} and DI_{10} metrics to analyze paper impact. Previous results indicate that many papers receive few citations or exhibit low disruption. We select papers with only one citation or a DI_{10} of zero. In Figure 10, the red histogram shows that most papers with a single citation have DI_{10} values near zero. However, over four hundred papers exhibit extremely high DI_{10} values in the close network. Similarly, the blue histogram indicates that many papers have an I_{10} less than 5, with the rest being outliers.

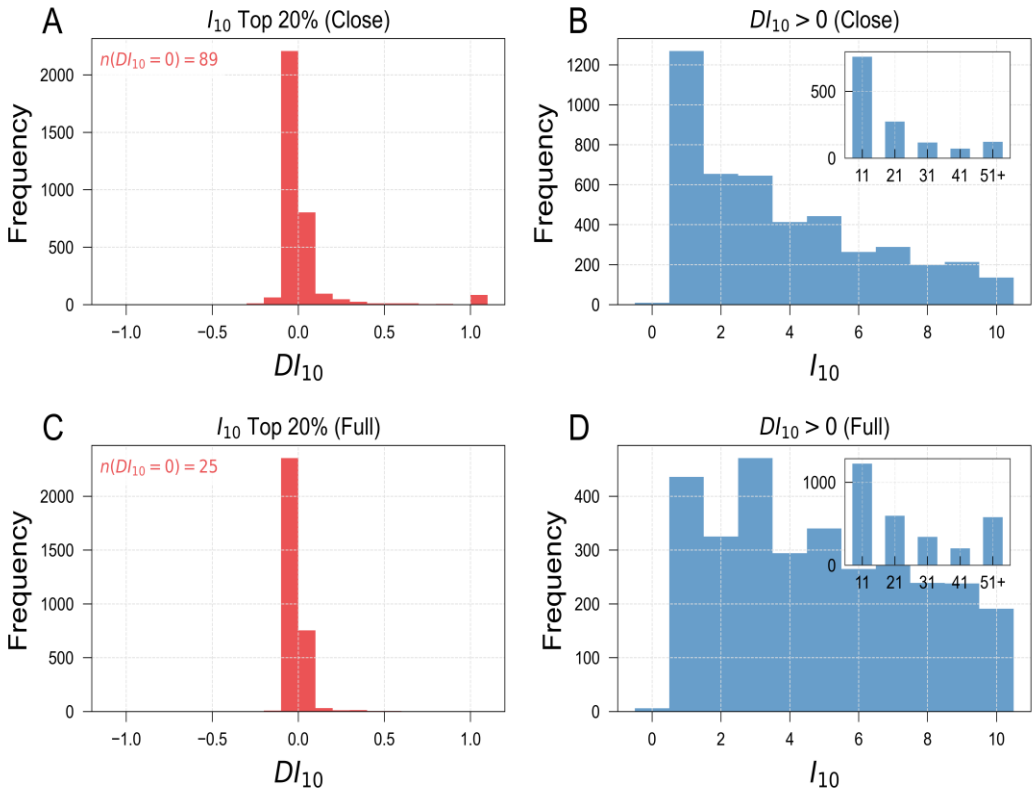


Figure 11. Distribution of DI_{10} and I_{10} values for papers with I_{10} in the top 20% or $DI_{10} > 0$.

We then examine papers with high impact. The red plot illustrates the DI_{10} distribution for papers in the top 20% of the I_{10} (thresholds: close = 14, full = 40). Most papers have DI_{10} values clustered around zero, with a sizable proportion below 0. We also analyze the papers with $DI_{10} > 0$, which typically rank in the lower 80% of the I_{10}. Table 5 further demonstrates this negative correlation. It is insignificant when the threshold is Top 1% and 5%. Figure 12 shows that a few outliers have both high impact and disruption.

Table 5. Negative correlation between I_{10} and DI_{10}.

Network	*Range (Top %)*	*Threshold*	*Sample*	*Correlation*	*p-value*
Close	1%	89	159	-0.162	p=.041
	5%	39	809	-0.115	p<.01
	10%	25	1653	-0.155	p<.001
	20%	14	3375	-0.170	p<.001
	100%	0	15701	-0.04	p<.001
Full	1%	255	159	-0.105	p=.187
	5%	107	800	-0.067	p=.059
	10%	69	1587	-0.12	p<.001
	20%	40	3195	-0.16	p<.001
	100%	0	15701	-0.298	p<.001

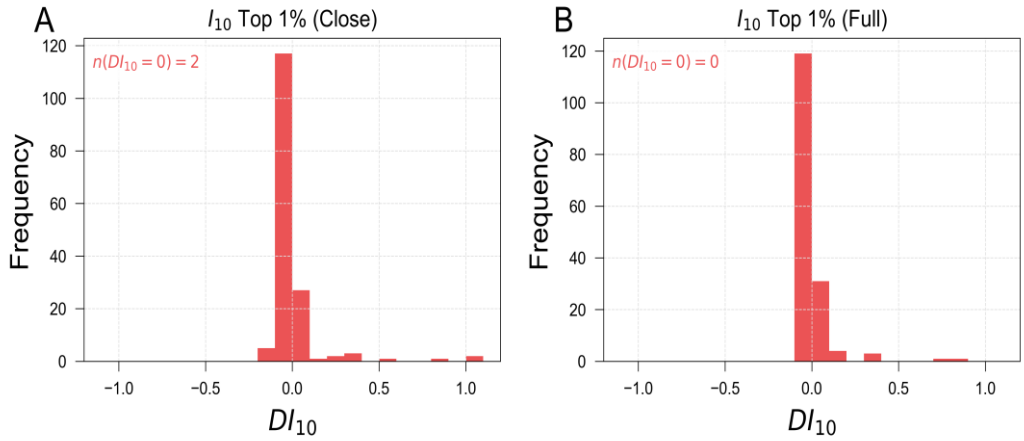


Figure 12. Distribution of DI_{10} values for papers with I_{10} in the top 1%.

Main Paths

We utilize Pajek to obtain five main paths with SPLC as the traversal count indicator and different selection methods. The main paths overlap and include 147 papers in total. Table 6 provides an overview. Diversity appears in the local forward path.

Table 6. Overview to main paths.

Main Paths	*Parameter*	*Nodes*	*Unique*
Global Standard	/	79	0
Local Backward	Tolerance=0.2	83	0
Local Forward	Tolerance=0.2	100	10
Local Key-route	Paths=1-20	130	3
Global Key-route	Paths=1-20	93	2

We merge the main paths for analysis. Table 7 shows the topics in different periods. From 1961 to 1983, early studies explored scientists' resistance to discoveries and

Matthew's effect on science. Co-citation analysis stood out in 1973 and ignited subsequent research in the 1980s. In 1991-2007, scholars discussed the journal's impact and research trends in the specific discipline. The third period enriched the knowledge in evaluating citation and journal impact. New indicators like success index and t-factor introduced new informetrics models. In the next period, scientists turned to bibliographic databases. They compared Scopus and WoS to analyze the data quality. Discussions on open platforms like Microsoft Academic Graph and Open Citations were also remarkable. We do not mention the last period because the relevant papers are not representative. In other words, they may not reflect the leading development of scientometrics in the last two years. A probable reason is the limitation of the MPA method itself. It relies on a sufficient citation window to determine the appropriate papers that appear on the main paths.

Table 7. Topics in the different periods of the main paths.

Period	*Main Topic*	*Count*
1961-1987	Co-citation analysis	31
1991-2007	Empirical studies with bibliometrics methods	18
2008-2016	Evaluation metrics	43
2016-2022	Bibliographic database	38

Figure 13 presents the DI value of papers along the main paths. Since only 86 papers appeared before 2015, we include DI_5 for papers published up to 2019, enabling a more comprehensive discussion. The final dataset comprises 118 papers. DI_{2024y} represents the most recent DI value. Each color block corresponds to a single paper, arranged chronologically with ten papers per row. Red indicates $DI > 0$, blue represents $DI < 0$, white denotes $DI = 0$, and grey signifies the absence of a DI value for the paper. The results show that most papers have $DI < 0$, while papers with $DI > 0$ cluster in the earlier years.

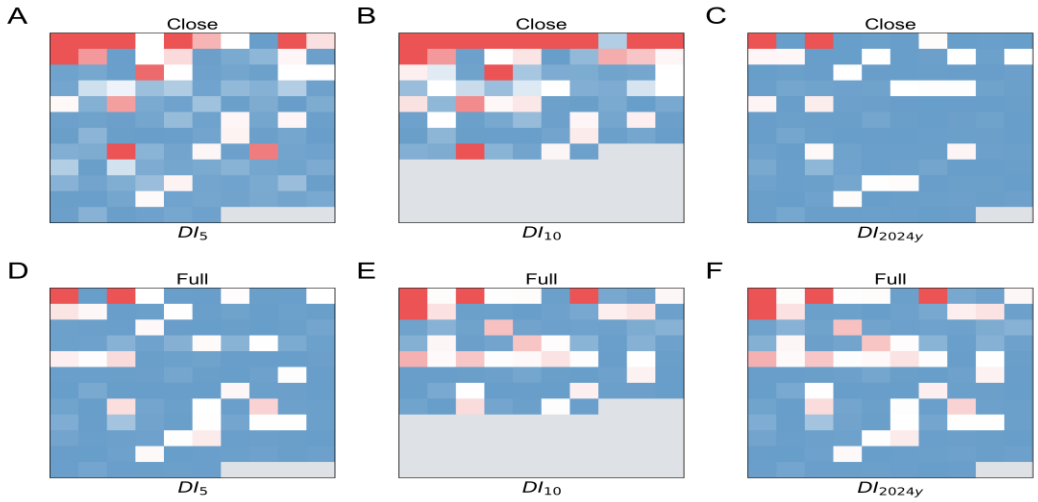


Figure 13. Distribution of DI values in different formats.

Additionally, we compare each paper with others published in the same year. Figure 14 demonstrates that all papers on the main paths exhibit DI_5 values higher than the annual average and median. However, this trend reverses significantly in DI_{10}. Regarding I_5 and I_{10}, papers on the main paths perform well within the close network but do not show a distinct advantage in the full network. One explanation is that some papers outside the main paths contribute to other fields.

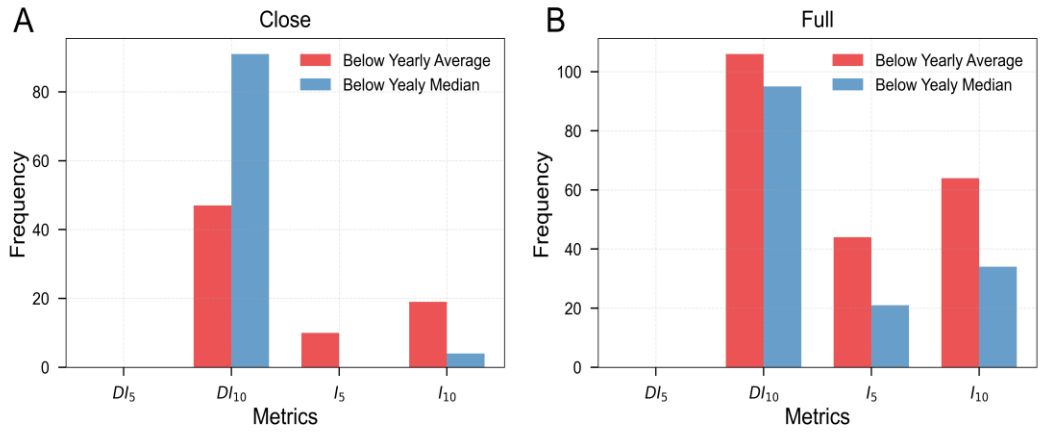


Figure 14. Comparison between papers on the main paths and others published in the same year.

The decline in values from DI_5 to DI_{10} catches further attention. Table 8 highlights this trend. Within the close network, all papers display a consistent decrease, while in the full network, some papers maintain higher DI values even 10 years after publication. The probable reason is that researchers from other fields adopt knowledge from scientometrics.

Table 8. Distribution of papers on the main paths with different relations on DI_5 and DI_{10}.

Relation	*Close*	*Full*
$DI_{10} < DI_5$	86	54
$DI_{10} = DI_5$	0	1
$DI_{10} > DI_5$	0	31

The papers on the main paths represent only a tiny fraction of the FPs. To broaden the scope of our analysis, we employ SPX, which measures a paper's contribution to knowledge flow within the citation network and reflects its indirect impact. Table 9 reports the Spearman correlation between SPX, DI, and I. To account for temporal variations, we apply different time windows, resulting in three groups of papers. The findings are significant and robust across the two networks, indicating a negative correlation between SPX and DI, while SPX shows a positive correlation with I.

Table 9. Spearman correlation between SPX, DI, and I.

Group	*Sample*	*Variable*	*Close*	*Full*
1955-2014	12,197	DI_5	-0.103***	-0.244***
		DI_{10}	-0.132***	-0.246***
		DI_{2024y}	-0.147***	-0.216***
		I_5	0.524***	0.356***
		I_{10}	0.512***	0.353***
		I_{2024y}	0.568***	0.438***
1955-2019	20,338	DI_5	-0.176***	-0.252***
		DI_{2024y}	-0.199***	-0.234***
		I_5	0.507***	0.367***
		I_{2024y}	0.540***	0.420***
1955-2024	32,042	DI_{2024y}	-0.210***	-0.223***
		I_{2024y}	0.200***	0.200***

*** $p < .001$

We further analyze papers with $DI > 0$ to explore the relationship between disruption and main path membership. Table 10 reveals that the negative correlation remains statistically significant.

Table 10. Negative correlation between SPX and DI.

Variable	*Time Span*	*Close*	*Sample*	*Full*	*Sample*
DI_5	1955-2019	-0.183***	6084	-0.144***	6901
DI_{10}	1955-2014	-0.144***	4973	-0.109***	4974
DI_{2024y}	1955-2024	-0.232***	9967	-0.190***	12511

*** $p < .001$

We employ the Monte Carlo simulation method to validate this observation, randomly assigning the "main path member" label while keeping the publication year constant. This approach allows us to simulate expected values under an unbiased condition. Figure 15 illustrates a consistent trend across all disruption metrics (DI_5, DI_{10}, and DI_{2024y}): the participation rate of highly disruptive papers in the main paths is consistently lower than the random baseline.

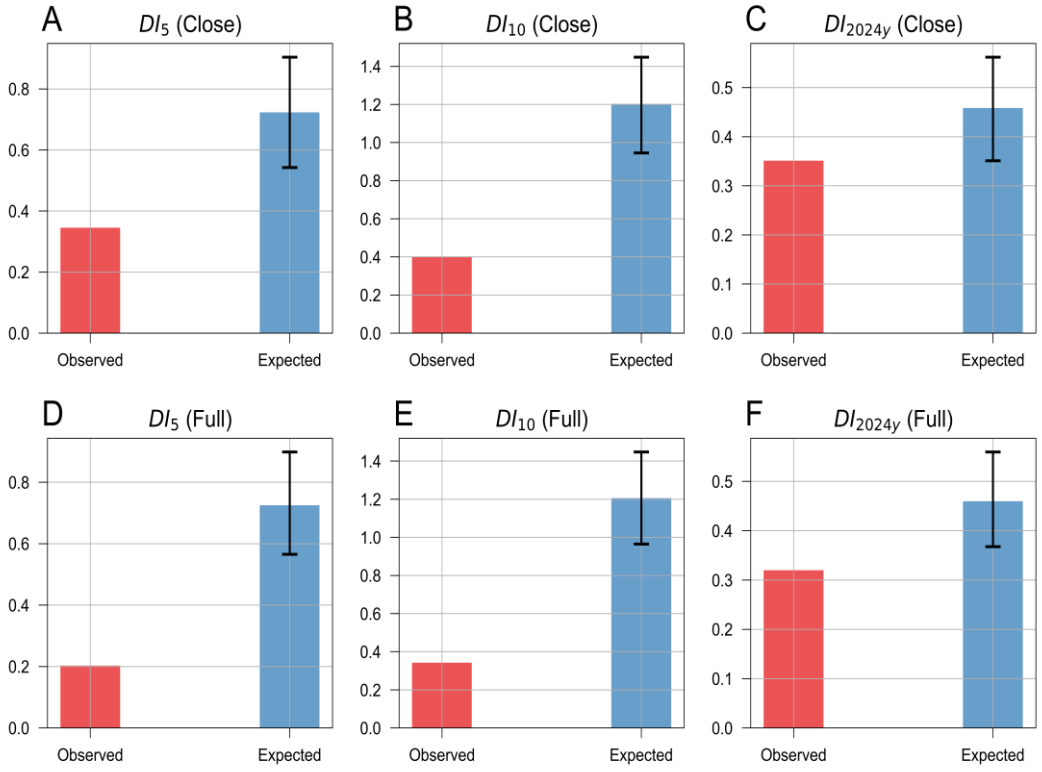


Figure 15. Participation of papers with $DI > 0$ on the main paths in the two situations.

Table 11. Statistic results for the validation experiment.

Variable	*Time Span*	*Close*	*OR (95%CI)*	*Full*	*OR (95%CI)*
DI_5	1955-2019	p<.001	0.388	p<.001	0.203
DI_{10}	1955-2014	p<.001	0.225	p<.001	0.187
DI_{2024y}	1955-2024	p<.1	0.691	p<.01	0.582

Additionally, the close and full networks exhibit similar patterns, suggesting that the observed results are independent of the citation network construction strategy. This trend demonstrates robustness across different network configurations. Table 11 provides detailed statistical evidence, showing that, except DI_{2024y} ($p = .06$ in the close network and $p = .002$ in the full network), the p-values for all other metrics are below 0.001.

In summary, we conclude that disruptive papers are significantly less likely to appear on the main paths.

Discussion and Conclusion

We analyze two metrics, I and DI, to examine papers in the scientometrics, with a particular emphasis on those situated on the main paths. Overall, papers on the main paths tend to exhibit lower disruption and demonstrate stronger consolidative tendencies over time. A comparative analysis with papers published in the same year reveals that this downward trend in disruption is significant. At the same time, these papers consistently show higher I values. However, their advantage in I diminishes when considering the attention from other fields.

Let us review the formula of the disruption:

$$DI = \frac{n_i - n_j}{n_i + n_j + n_k}$$

Here, j-type papers, which cite both the FP and the references of the FP, contribute directly to a negative impact on disruption. We hypothesize that the coupling relationships among main path members play a key role in reducing disruption. Figure 16 provides evidence for this hypothesis. For each j-type descendant of a paper, we identified b-type papers that cite both the FP and other members of the main paths. We then calculate the proportion of b-type papers within the j-type set. The results indicate that b-type papers significantly increase the number of j-type papers, thereby reducing DI values.

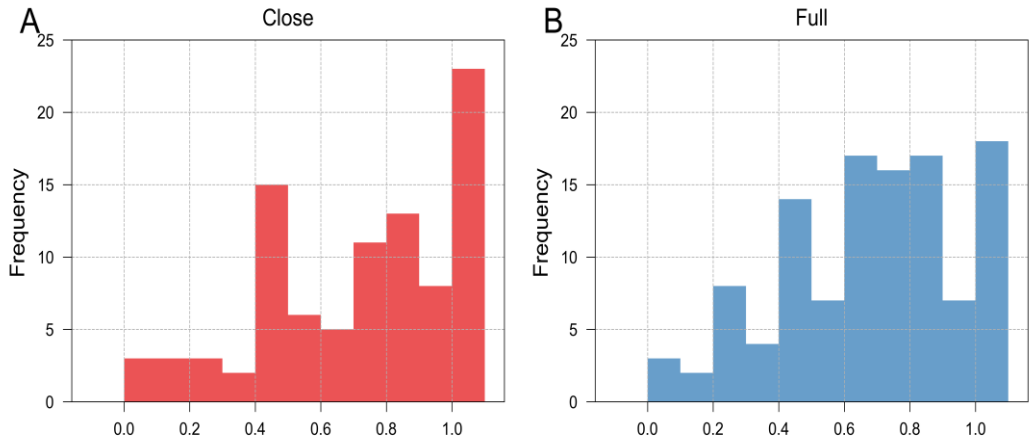


Figure 16. Proportion of b-type within j-type papers for the members of the main paths.

We introduce the SPX to examine the relationship between direct impact, indirect impact, and disruption. Our findings show that indirect impact is positively correlated with direct impact, while both negatively correlate with disruption.

The top 1% of highly influential papers form a distinct group. The sample size ($n \in (100,350)$, depending on the time window) influences the robustness and significance of these correlations. For example, outliers with high influence and high disruption weaken the observed negative correlation. Similarly, some papers with exceptionally high impact fail to achieve high SPX values. This discrepancy arises

because network structure is critical in determining SPX values. Notably, the correlations regain statistical significance when we set the threshold from the top 1% to the top 5%. Future studies could investigate their topics and citation patterns to provide deeper insights into their unique characteristics.

On the other hand, we specifically focus on FPs with $DI > 0$, where their SPX values demonstrate a consistently stable negative correlation with DI. Statistical analyses further indicate that disruptive papers are less likely to be part of the main paths.

Our study provides a multidimensional evaluation framework. It can bring a more comprehensive understanding of how papers contribute to scientific progress. Future research could further investigate it across different disciplines.

In addition, we focus on analyzing papers along the main path. The main path mechanism prioritizes and amplifies conventional scientific achievements, creating a "highway" for knowledge diffusion. In contrast, disruptive papers are more likely to spread through smaller, less prominent paths, suggesting a divergence in the dissemination patterns of traditional and disruptive contributions.

This study also offers two practical recommendations. First, we propose giving greater attention to non-mainstream breakthroughs when assessing the impact of papers, as these contributions may represent emerging or unconventional advancements. Second, main path analysis may not be the suitable tool for identifying disruptive technological frontiers, given its inherent focus on established knowledge trajectories.

There still exist certain limitations. It is difficult to reduce noise in the dataset like incorrect citation relationships and papers that do not belongs to scientometrics, which may affect identifying the main paths. Besides, we only adopt SPLC as the link traversal algorithms, introducing advanced approaches could help optimize the results. Additionally, the SPX indicator covers only about 90% of the nodes, as calculating SPLC values in Pajek requires selecting the largest subnetwork. Future research could explore methods to address this constraint and ensure more comprehensive coverage. Finally, we do not disclose the difference in citation patterns between main path members and disruptive papers in detail. Case study may bring more insight into how the two kinds of papers contribute the scientific progress in scientometrics.

Acknowledgments

We appreciate reviewers' helpful suggestions. We acknowledge Ruitao Zhang, Xiao Tan, and Jiting Yi for their assistance with data collection. We also thank OpenAlex for providing powerful APIs. Prof. Rongying Zhao offers suggestions. Discussion with Elysia inspires us.

References

Azoulay, P. (2019). Small research teams 'disrupt' science more radically than large ones. *Nature, 566*(7744), 330–332.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century—A review. *Journal of Informetrics, 2*(1), 1–52.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. arXiv.

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020a). Disruptive papers published in scientometrics: Meaningful results by using an improved variant of the disruption index originally proposed by wu, wang, and evans (2019). *Scientometrics*, *123*(2), 1149–1155.

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020b). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, *1*(3), 1242–1259.

Bornmann, L., & Tekles, A. (2019a). Disruption index depends on length of citation window. *El Profesional De La información*, *28*(2).

Bornmann, L., & Tekles, A. (2019b). Disruptive papers published in scientometrics. *Scientometrics*, *120*(1), 331–336.

Bornmann, L., & Tekles, A. (2021). Convergent validity of several indicators measuring disruptiveness with milestone assignments to physics papers by experts. *Journal of Informetrics*, *15*(3), 101159.

Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. *Quantitative Science Studies*, *2*(1), 155–183.

Chen, J., Shao, D., & Fan, S. (2021). Destabilization and consolidation: Conceptualizing, measuring, and validating the dual characteristics of technology. *Research Policy*, *50*(1), 104115.

Chen, L., Xu, S., Zhu, L., Zhang, J., Xu, H., & Yang, G. (2022). A semantic main path analysis method to identify multiple developmental trajectories. *Journal of Informetrics*, *16*(2), 101281.

Chen, S., Guo, Y., Ding, A. S., & Song, Y. (2024). Is interdisciplinarity more likely to produce novel or disruptive research? *Scientometrics*, *129*(5), 2615–2632.

Deng, N., & Zeng, A. (2023). Enhancing the robustness of the disruption metric against noise. *Scientometrics*, *128*(4), 2419–2428.

Everton, S. F., de Nooy, W., Mrvar, A., & Batagelj, V. (2018). *Exploratory social network analysis with pajek: Revised and expanded edition for updated software* (3rd). Cambridge: Cambridge University Press.

Filippin, F. (2021). Do main paths reflect technological trajectories? Applying main path analysis to the semiconductor manufacturing industry. *Scientometrics*, *126*(8), 6443–6477.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791–817.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Ho, M. H.-C., Liu, J. S., & Chang, K. C.-T. (2017). To include or not: The role of review papers in citation-based analysis. *Scientometrics*, *110*(1), 65–76.

Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, *11*(1), 39–63.

Jiang, X., Zhu, X., & Chen, J. (2020). Main path analysis on cyclic citation networks. *Journal of the Association for Information Science and Technology*, *71*(5), 578–595.

Jiang, Y., & Liu, X. (2023). A construction and empirical research of the journal disruption index based on open citation data. *Scientometrics*, *128*(7), 3935–3958.

Kuan, C.-H. (2023). Does main path analysis prefer longer paths? *Scientometrics*, *128*(1), 841–851.

Kuan, C.-H., & Liao, S.-Y. (2024). Assessing main paths by uncovering their coverage with key-node path search. *Scientometrics*, *129*(11), 6629–6657.

Leahey, E., Lee, J., & Funk, R. J. (2023). What Types of Novelty Are Most Disruptive? *American Sociological Review*, *88*(3), 562–597.

Leibel, C., & Bornmann, L. (2024). What do we know about the disruption index in scientometrics? An overview of the literature. *Scientometrics*, *129*(1), 601–639.

Leydesdorff, L., & Bornmann, L. (2021). Disruption indices and their calculation using web-of-science data: Indicators of historical developments or evolutionary dynamics? *Journal of Informetrics*, *15*(4), 101219.

Leydesdorff, L., Tekles, A., & Bornmann, L. (2021). A proposal to revise the disruption index. *El Profesional De La información*, e300121.

Li, H., Tessone, C. J., & Zeng, A. (2024). Productive scientists are associated with lower disruption in scientific publishing. *Proceedings of the National Academy of Sciences*, *121*(21), e2322462121.

Li, J., & Chen, J. (2022). Measuring destabilization and consolidation in scientific knowledge evolution. *Scientometrics*, *127*(10), 5819–5839.

Liang, G., Lou, Y., & Hou, H. (2022). Revisiting the disruptive index: Evidence from the nobel prize-winning articles. *Scientometrics*, *127*(10), 5721–5730.

Lin, Y., Evans, J. A., & Wu, L. (2022). New directions in science emerge from disconnection and discord. *Journal of Informetrics*, *16*(1), 101234.

Lin, Y., Frey, C. B., & Wu, L. (2023). Remote collaboration fuses fewer breakthrough ideas. *Nature*, *623*(7989), 987–991.

Liu, J. S., & Kuan, C. (2016). A new approach for main path analysis: Decay in knowledge diffusion. *Journal of the Association for Information Science and Technology*, *67*(2), 465–476.

Liu, J. S., & Lu, L. Y. Y. (2012). An integrated approach for main path analysis: Development of the hirsch index as an example. *Journal of the American Society for Information Science and Technology*, *63*(3), 528–542.

Liu, J. S., Lu, L. Y. Y., & Ho, M. H.-C. (2019). A few notes on main path analysis. *Scientometrics*, *119*(1), 379–391.

Liu, J. S., Lu, L. Y. Y., & Ho, M. H.-C. (2020). A note on choosing traversal counts in main path analysis. *Scientometrics*, *124*(1), 783–785.

Liu, X., Bu, Y., Li, M., & Li, J. (2024). Monodisciplinary collaboration disrupts science more than multidisciplinary collaboration. *Journal of the Association for Information Science and Technology*, *75*(1), 59–78.

Lyu, D., Gong, K., Ruan, X., Cheng, Y., & Li, J. (2021). Does research collaboration influence the "disruption" of articles? Evidence from neurosciences. *Scientometrics*, *126*(1), 287–303.

Ma, V. C., & Liu, J. S. (2016). Exploring the research fronts and main paths of literature: A case study of shareholder activism research. *Scientometrics*, *109*(1), 33–52.

Macher, J. T., Rutzer, C., & Weder, R. (2024). Is there a secular decline in disruptive patents? Correcting for measurement bias. *Research Policy*, *53*(5), 104992.

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*(7942), 138–144.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv.

Ruan, X., Lyu, D., Gong, K., Cheng, Y., & Li, J. (2021). Rethinking the disruption index as a measure of scientific and technological advances. *Technological Forecasting and Social Change*, *172*, 121071.

Sheng, L., Lyu, D., Ruan, X., Shen, H., & Cheng, Y. (2023). The association between prior knowledge and the disruption of an article. *Scientometrics*, *128*(8), 4731–4751.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Wang, R., Zhou, Y., & Zeng, A. (2023). Evaluating scientists by citation and disruption of their representative works. *Scientometrics*, *128*(3), 1689–1710.

Wang, Y., Li, N., Zhang, B., Huang, Q., Wu, J., & Wang, Y. (2023). The effect of structural holes on producing novel and disruptive research in physics. *Scientometrics*, *128*(3), 1801–1823.

Wang, Z., Zhang, H., Chen, J., & Chen, H. (2024). An effective framework for measuring the novelty of scientific articles through integrated topic modeling and cloud model. *Journal of Informetrics*, *18*(4), 101587.

Wei, C., Li, J., & Shi, D. (2023). Quantifying revolutionary discoveries: Evidence from nobel prize-winning papers. *Information Processing & Management*, *60*(3), 103252.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382.

Yan, Z., & Fan, K. (2024a). An integrated indicator for evaluating scientific papers: Considering academic impact and novelty. *Scientometrics*, *129*(11), 6909–6929.

Yan, Z., & Fan, K. (2024b). A multi-entity reinforced main path analysis: Heterogeneous network embedding considering knowledge proximity. *Journal of Informetrics*, *18*(4), 101593.

Yang, A. J., Gong, H., Wang, Y., Zhang, C., & Deng, S. (2024). Rescaling the disruption index reveals the universality of disruption distributions in science. *Scientometrics*, *129*(1), 561–580.

Yang, A. J., Hu, H., Zhao, Y., Wang, H., & Deng, S. (2023). From consolidation to disruption: A novel way to measure the impact of scientists and identify laureates. *Information Processing & Management*, *60*(5), 103420.

Yang, A. J., Yan, X., Hu, H., Hu, H., Kong, J., & Deng, S. (2024). Are disruptive papers more likely to impact technology and society? *Journal of the Association for Information Science and Technology*, asi.24947.

Yang, S., & Kim, S. Y. (2023). Knowledge-integrated research is more disruptive when supported by homogeneous funding sources: A case of US federally funded research in biomedical and life sciences. *Scientometrics*, *128*(6), 3257–3282.

Zeng, A., Fan, Y., Di, Z., Wang, Y., & Havlin, S. (2021). Fresh teams are associated with original and multidisciplinary research. *Nature Human Behaviour*, *5*(10), 1314–1322.

Zhao, Y., Wang, Y., Zhang, H., Kim, D., Lu, C., Zhu, Y., & Zhang, C. (2024). Do more heads imply better performance? An empirical study of team thought leaders' impact on scientific team performance. *Information Processing & Management*, *61*(4), 103757.

Zhou, Y., Xu, X.-L., Yang, X.-H., & Li, Q. (2022). The influence of disruption on evaluating the scientific significance of papers. *Scientometrics*, *127*(10), 5931–5945.

https://doi.org/10.51408/issi2025_090

Quantifying the Political Attributes of Technology for Potential Bottleneck Technologies Identification: Evidence from Chinese Integrated Circuits Industry

Tao Zhiyu[1], Liang Shuang[2], Li Hanxi[3], Liu Yajing[4]

[1] *taozhiyu@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[2] *liangs1998@126.com*
Department of Information Management, Peking University, Beijing (China)

[3] *lihanxi24@mails.ucas.ac.cn*
School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[4] *yajingliu@cau.edu.cn*
College of Humanities and Development Study, China Agricultural University, Beijing (China)

Abstract

Technological innovations are becoming increasingly competitive among nations, as countries strive to gain a technological advantage to safeguard their national interests. This competition leads to technology suppression, supply disruption, and export controls, which can undermine the integrity of supply chains. Technologies supply disrupted by export controls from collaborating countries are referred to as bottleneck technologies, posing significant threats to national security. These technologies shall be identified promptly to inform effective technology and diplomacy policymaking. Existing studies have focused on the quantity and quality gaps or topic strength gaps of technologies, emphasizing their technological attributes. However, political attributes, particularly those driven by political competition, have received insufficient attention. We argue that bottleneck technologies are not only technological products but also political products, shaped by both technological and political factors. This paper introduces the concept of 'technological political distance' to identify bottleneck technologies, characterized by a country's subjective motivation to create a 'control.' By analyzing citation networks and calculating indices like PageRank as “be able to control”, we identify highly cited patents in key technology areas as 'worthwhile to control' in terms of value. Empirical research in the field of integrated circuits shows that China faces high risks in foundational semiconductor technologies, circuit integration methods, material science, and manufacturing processes, while the risks in sensor, imaging, and signal transmission technologies are relatively low.

Introduction

Science and technology (S&T) innovation has become a critical arena of national competition, with countries vying for emerging and advanced technologies to secure global competitive advantage(Schmid et al., 2025). This intense rivalry not only heightens technological competition but also disrupts international technological collaboration, posing significant threats to national security(Luo, 2022; Sun, 2019; Vivoda, 2023). Consequently, it is crucial to identify potential bottleneck

technologies and assess the associated risks, so that policymakers can both leverage the dividends of global collaboration and safeguard S&T security. Drawing on historical instances of international technology competition—particularly the U.S.-China rivalry, this paper argues that bottleneck technologies are not merely a technical concern but also a political one, exhibiting intertwined attributes of technology and politics. We extend the literature on technology identification and bottleneck technologies (Guoxiong et al., 2021; Haiqiu et al., 2023; Jin et al., 2020; Zhiwei et al., 2021) by conceptualizing bottleneck technologies as those characterized by (1) the willingness to impose technology controls, (2) the capacity to impose such controls, and (3) the strategic value that motivates these controls. To quantitatively evaluate these attributes, we incorporate a Political Distance (PD) index—calculated from large-scale United Nations (U.N.) voting data—to quantify geopolitical risks encountered by technologies and construct a citation network to represent the overall technology system. We then apply the PageRank algorithm to identify key technologies which play key roles in maintain the function and integrity of the technology system, whose removals may cause the system dismantling and technology dysfunction. Combining patent-based and topic-based analyses, we propose that those bottleneck technologies are controlled by competitors who both desire and are able to halt supply to China, and which China cannot rapidly reproduce. An empirical study on integrated circuits demonstrates that China is highly vulnerable in foundational areas such as semiconductor devices, circuit integration methods, material science, and manufacturing processes, yet faces relatively lower risks in sensor technology, imaging technology, signal transmission, and other applications. These findings are validated by expert assessments and the U.S. technology control list, highlighting the practical utility of this method.

Methodology and Research Design

This study introduces a novel metric, **Technology Political Distance (TPD),** to quantify the political risks associated with various technologies. The metric is derived from extensive voting data sourced from the United Nations. Additionally, this research incorporates PageRank-based algorithms to identify technologies that are central to the overall technological ecosystem. By combining these approaches, the study highlights key bottleneck technologies at both the patent and topic levels. The proposed research framework is visually represented in Figure 1.

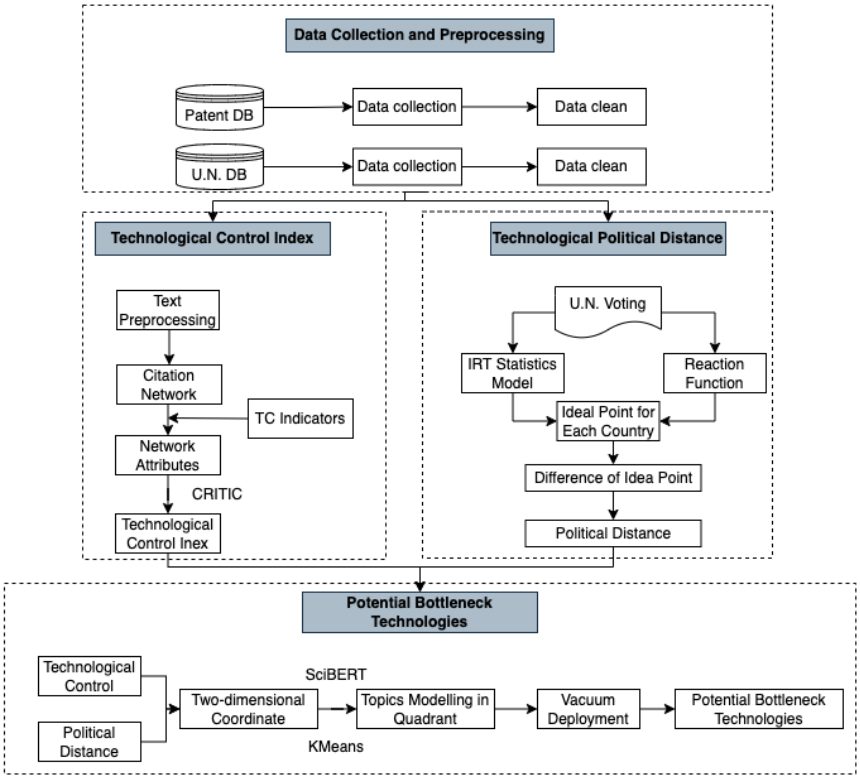


Figure 1. Research framework.

Quantifying the Political Attributes: Technology Political Distance

This paper introduces the concept of political distance to analyse the potential effects of international collaboration across different countries. Drawing on the idea proposed by Bailey et al. (2016), political distance is characterized using discrepancies in countries' voting behaviors at the United Nations on various issues. These voting differences act as proxies for the political distance between nations. A larger voting disparity between two countries typically reflects divergent national interests, increasing the likelihood of rivalry. In contrast, a smaller voting difference indicates closer alignment in interests, suggesting a higher probability of these countries being allies or partners. To enhance the accuracy of the political distance measure, we employ the Item Response Theory (IRT) statistical model, which constructs annual scale data representing each country's "ideal point"—a binary metric indicating the shifting similarity in political preferences between two countries. The IRT model, traditionally used to describe the relationship between a subject's latent traits (such as abilities) and their responses to test items, is adapted here to estimate the ideal point, which reflects a country's foreign policy orientation. This methodology provides a more nuanced and robust framework for measuring political distance in international relations.

$$\Pr(Y_{itv} = K) = \Phi(r_{kv} - \beta_v \theta_{itv}) - \Phi\left(r_{k-1,v} - \beta_v \theta_{itv}\right) \quad (1)$$

In the above equation, the left-hand side represents the probability distribution of country i's choice of approval (k=1), abstention (k=2), and negation (k=3) in the v-th vote, which can be obtained by observing the voting behavior. Where β represents the differentiation parameter of the item, r represents the difficulty parameter of the item, and θ represents the ideal point of the measured ability or trait, the posterior expectations of the parameters β, r, and θ can be estimated using Bayesian estimation with the help of MCMC (Markov Chain Monte Carlo) algorithm.

Further, following Davis et al. (2019), the absolute difference between the ideal points of China and its partner countries is employed as a proxy for bilateral political distance. This metric specifically quantifies the degree of divergence between China's foreign policy orientation and that of its trading partners, thereby providing an indicator of the political relationship between the two nations. This approach offers a more precise measure compared to traditional indices such as the voting similarity index, the affinity index, and the "S" index. So, we employ the divergence of ideal point distance to quantify the political distance between countries. By following these steps, the political distance between China and other countries can be calculated.

Since a single patent may belong to multiple patent families registered across different countries, it is essential to consider the patent family structure. We argue that expanding a patent family across multiple nations generates substantial technology spillover effects in the current market (Frakes & Wasserman, 2021; Lee, 2021; Taichen et al., 2022). This expansion can accelerate technology transfer and foster local technological development (Xue, 2022), driving technological advancement and industry upgrading. Building on this, we hypothesize that when countries with significant political distance from China register patents either within China or in countries with close technological proximity to China, the resulting technology spillover can stimulate local technological growth and upgrading. This, in turn, reduces the likelihood of these technologies becoming bottlenecks for China. On the other hand, if countries with considerable political distance from China register patents in other nations that also maintain substantial political distance from China, these countries are more likely to form technological alliances and establish barriers, which could restrict China’s access to these technologies. Based on this framework, we define TPD as the average political distance between the countries where the patent family is registered and China, denoted as:

$$TPD_n = \frac{\sum_{i=1}^{m} PD_i}{m} \tag{2}$$

Technology Control Capability

Motivated by technology system theory as proposed by Arthur (2009), we conceptualize the entire technological landscape as a complex system. Building on prior studies that employ complex networks to model such systems (Han et al., 2021), we construct a citation network to capture the interconnections and structural composition of the technology ecosystem. In the context of technology competition, the control over certain key technologies has been observed to disrupt the proper

functioning of an entire technological field. To further explore this phenomenon, we introduce the concept of network dismantling, which involves the strategic removal of specific nodes (i.e., technologies) to fragment the citation network and induce dysfunction within the broader technology field (Fan et al., 2020). The identified nodes represent potentially risky technologies, whose removal could critically impair technological continuity and development.

Building on this concept, we introduce a network-based algorithmic approach to identify critical technologies—those essential to maintaining the integrity of the technology system. Given that different algorithms assess node importance from varying perspectives, we integrate multiple algorithms to create a complementary framework for identifying key technologies more effectively. To achieve this, we employ degree centrality (DC), betweenness centrality (BC), and structural hole (SH) analysis (S. Burt, 1992), along with HITS and PageRank (PR) (Tongliang et al., 2023). These measures collectively capture different dimensions of a technology's influence within the network: (1) Degree centrality (DC) identifies technologies with the highest number of direct connections. (2) Betweenness centrality (BC) detects technologies that serve as critical bridges between different subfields. (3) Structural hole (SH) highlights technologies that control access to otherwise disconnected technological domains(S.Burt, 1992). (4) HITS (Hyperlink-Induced Topic Search) distinguish between hub technologies (those that connect to many authoritative technologies) and authority technologies (those that are referenced by influential hubs). (5) PageRank (PR) assigns importance based on the recursive influence of a technology within the citation network (Tongliang et al., 2023). By leveraging this multi-perspective approach, we enhance the robustness of our analysis, ensuring a more comprehensive identification of crucial technologies within the system.

To comprehensively assess the weight of each indicator, we employ the Criteria Importance Through Intercriteria Correlation (CRITIC) algorithm, a well-established method for determining indicator importance (Danae et al., 1995). The CRITIC algorithm evaluates the significance of each indicator by analyzing both its comparative strength and its degree of conflict with other indicators. Through this approach, the weight of each indicator is systematically determined based on its intrinsic information content and its correlation with other indicators. The calculation of indicator weights follows the methodology outlined below:

$$W_j = \delta_j \sum_{j=1}^{n} \left(1 - R_{kj}\right), j \neq k, j = 1,2, \ldots, n \tag{3}$$

W_j denotes the weight for indicator j, and R_{kj} represents the correlation between the k-th indicator and the j-th indicator.

Based on the weight and value of each indicator, the TC for each technology can be calculated by:

$$TC_i = W_{DC_norm} * DC_norm(i) + W_{BC_norm} * BC_norm(i) + W_{SH_norm} * SH_norm(i) + W_{HITS_norm} * HITS_norm(i) + W_{PR_norm} * PR_norm(i) \tag{4}$$

The TC_i denote the technology control capability of i th technology, and W_{DC_norm}, W_{BC_norm}, W_{SH_norm}, W_{HITS_norm}, W_{PR_norm} denote the weight of DC, BC, SH, HITS, PR respectively which have been normalized and the weight is calculated by CRITIC.

Technologies classification based on dual perspective of politics and technology

According to the dual properties of technologies in TC and PD perspectives, we categorize technologies into four types as Type A (high TC and high PD), indicating risky technologies due to those highly-impact technologies which are important to technology system are held by rival countries who have great PD with our country. Type B (high TC and low PD) is friendly sophisticated technologies held by our country and friendly countries. Type C and Type D are low-impact technologies, which exert limited impact on the technology system, so, whether those technologies held by our country, friendly countries or rivals will not significantly influence the normal operation of technology system, so, they are difficult to be the bottleneck technologies.

Based on this classification (Figure 2), those technologies exist in Type A but do not appear in Type B are those sophisticated technologies held by rivals but not held by us and our friends, which can be taken as highly risky potential technologies, which is our general idea on bottleneck technologies identification.

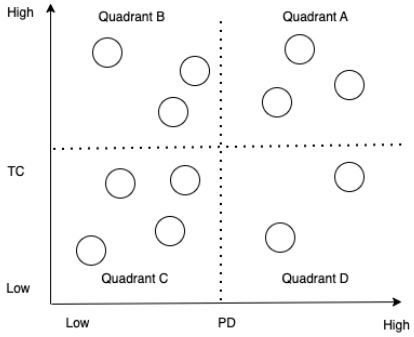


Figure 2. Four types of technologies classified by PD and TC.

Potential bottleneck technologies identification

Based on quantifying the PD and TC, we identify potential bottleneck technologies on patent and topic level respectively to complement the micro and macro information. In micro level, we propose the bottleneck index as K index which is defined as:

$$K_i = TC_i * TPD_i \tag{5}$$

K index describes whether those highly impact technologies are held by rival countries, to reflect the risk of be controlled in both technological and political perspective.

Furthermore, given that technology export control lists typically reference clusters of technologies rather than isolated patents, we conceptualize these clusters as "technology topics." To extract these topics, we first obtain the abstract text from each patent and employ SciBERT which is proposed by Beltagy et al. (2019) to convert the text into semantic vectors, ensuring that words with similar meanings are positioned closely in the semantic space. Next, we apply the K-means clustering algorithm to group semantically similar words, thereby forming coherent technology topics. Finally, we compare the semantic similarity between topics in Type A and Type B—using a threshold of 0.8 to indicate identical topics. Technologies associated with topics that appear in Type A but not in Type B are classified as potential bottleneck technologies, whereas those found in Type B but absent from Type A are identified as strategic advantage technologies that could inform the implementation of technology sanctions.

Empirical Study: Initial Results on Chinese Integrated Circuits Fields

Data Source and Preprocessing

Integrated circuits (IC) are at the heart of modern information technology and the electronics industry. As core technologies, they are pivotal for building national competitive advantages in the digital age and have become a central arena in the U.S.-China technology competition. Accurately identifying potential bottleneck technologies in the IC domain is therefore essential for maintaining national security. Furthermore, recent U.S. export controls on various IC technologies have intensified bottleneck effects. The methodology proposed in this study, which does not rely on pre-tested information and can be validated through an actual list of bottleneck technologies, offers timely insights into these challenges. For these reasons, the IC sector was selected for our empirical analysis. Patent data were retrieved from the Derwent Innovation Index (DII) database using the manual coding system developed by Derwent experts. We employed the retrieval formula "U13-*" on 30 November 2023, which returned a total of 290,743 patents. Recognizing that bottleneck technologies are often characterized by high-value patents—as reflected in their citation counts—we filtered the dataset to retain only those patents with at least five citations, resulting in a subset of 83,211 patents. Finally, comprehensive data cleaning and preprocessing procedures were applied to ensure the dataset's readiness for further analysis.

Political Distance Calculation

According to the Equation 1, we utilize the IRT reaction function to calculate the ideal point for each country respectively, and calculate the absolute difference of ideal point between each country pairs. Notably, for organizations such as the European Patent Office (EP) and the World Intellectual Property Organization (WO), we calculate their TPD as the average PD between China and the participating countries within each organization. Further we extract the country (organization) of each patent holder and have 32 countries in total, and list those 5 countries with largest and closest political distance from China as shown in Table 1:

Table 1. The five countries with the largest and closest political distance from China.

Country	PD
US	3.116
IL	2.952
GB	2.179
CA	2.082
FR	1.962
BR	0.318
MY	0.266
SG	0.242
ZA	0.203
IN	0.161

Technology Control Capability

Based on the patent citation network, we apply the five network-based algorithms to calculate the PR index and other indicators for each patent, and apply the Equation 3 for evaluating the weight for each indicator as shown in Table 4 and calculate the TC for each patent by Equation 4. We list the 5 patents which have the highest TC as shown in Table 3.

$$W_j = \delta_j \sum_{j=1}^{n} \left(1 - R_{kj}\right), j \neq k, j = 1,2, \dots, n \tag{5}$$

Table 2. Weight for each indicator calculated by CRITIC.

Indicator	Weight
ω_{DC}	0.0562
ω_{BC}	0.0371
ω_{HITS}	0.0265
ω_{PR}	0.0258
ω_{SH}	0.0543

Table 3. Patents with top 5 TC.

PN	TC
US2006007612-A1	0.133
WO9907000-A2	0.126
EP1746645-A2	0.118
US2007196982-A1	0.112
EP738010-A2	0.105
…	…

Potential Bottleneck Technologies Identification: patent and topic level

According to the definition and method for quantifying the bottleneck technologies (Equation 5), we first calculate the K index for each patent and list those patents with top 5 K index as shown in Table 4.

By reading the abstract of those five patents, we find that they are in the technology field of: (1) circuit design for protecting nonvolatile read-only memories; (2) programming methods for nonvolatile memory cells; (3) reverse read-programmed EEPROMs and ROMs; (4) process and structural optimization of nonvolatile memory arrays; (5) construction of imaging sensors.

Table 4. Patents with top 5 K Index.

Rank	PN	K Index
1	US2006007612-A1	0.133
2	WO9907000-A2	0.126
3	EP1746645-A2	0.118
4	US2007196982-A1	0.112
5	EP738010-A2	0.105
…	…	

To classify the technologies into four distinct categories, we use two threshold criteria: the median value of TPD and the 80th percentile of TC. These thresholds, indicated by the red dashed lines in Figure 3, divide the dataset into four quadrants, with each quadrant representing a unique category of technology.

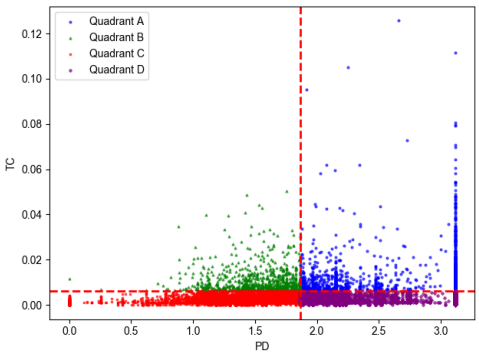


Figure 3. The distribution of four types of technologies.

To evaluate the topic distribution within Quadrant A (Type A technologies) and Quadrant B (Type B technologies), we employ a two-step approach. First, we use the SciBERT-Kmeans method to extract technology topics. However, since the number of topics for each technology type must be determined manually, we then apply Latent Dirichlet Allocation (LDA) for topic modeling to determine the proper number of topics. For each technology type, we calculate the coherence score to assess model quality and select the number of topics that yields the highest coherence

score. Based on this analysis, we define 14 topics for Type A technologies and 20 topics for Type B technologies. The resulting topic distributions are presented in Figure 4.

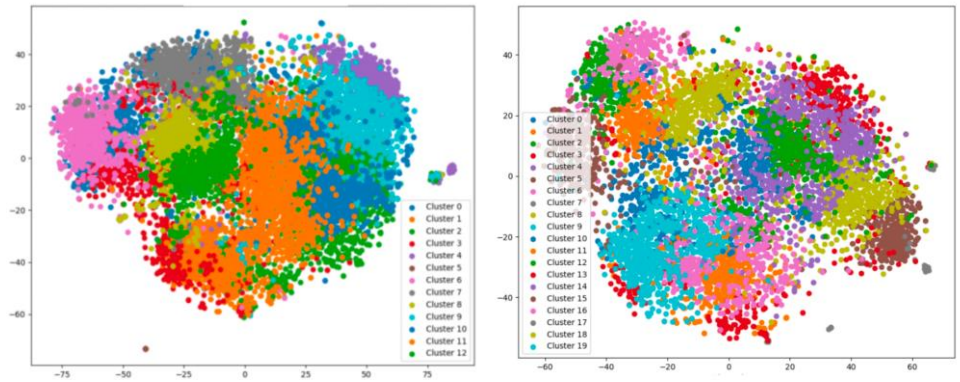


Figure 4. Topic distribution for Quadrant A(left) and Quadrant B(right).

Furthermore, we employ the Term Frequency-Inverse Document Frequency (TF-IDF) method to extract the top 30 keywords representing each topic, subsequently inviting domain experts to label each topic based on these keywords. Our analysis reveals that the topics in Quadrant A primarily pertain to semiconductor devices, circuit integration, logic devices, insulation technology, electrode engineering, imaging and sensing, logic circuits, electrode integration, oxidation technology, electrical signals, electrode dynamics, imaging integration, as well as insulation and electrode-related fields. This indicates that the technologies in Quadrant A predominantly focus on the manufacturing and design of semiconductor devices. Similarly, the topics in Quadrant B encompass areas such as imaging processors, insulated circuits, signal imaging, semiconductor devices, line transmission, storage arrays, selective thin films, photoelectric imaging, signal films, circuit components, voltage thin films, sensing imaging, insulated storage, imaging thin films, insulating films, sensing transistors, signal gates, semiconductor surface engineering, insulated circuits, and voltage equipment.

To compare the topic similarity between topics in two category, we apply the cosine similarity calculation on topics' semantic vector which can be found in Figure 5.

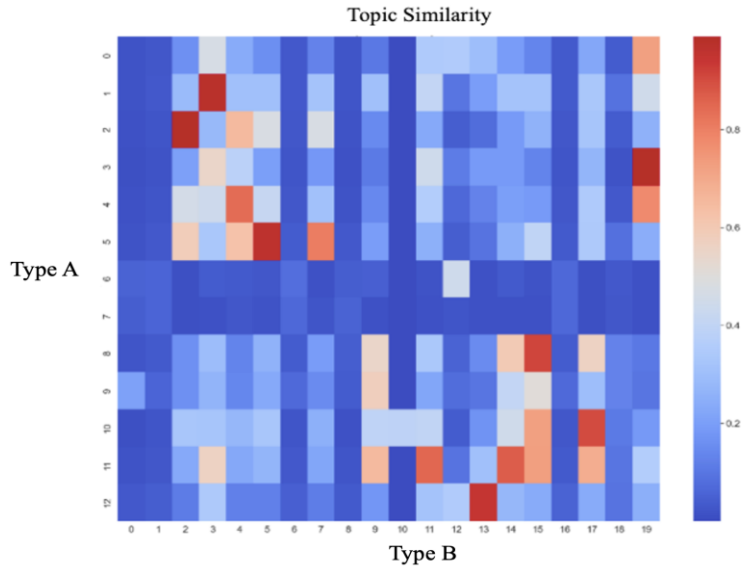


Figure 5. The topic similarity between Quadrant A(y axis, Type A) and Quadrant B(x axis, Type B).

As illustrated in Figure 5, our analysis reveals that in our (our country and friendly country with small TPD) Topic 0, 1, 6, and 7 (Type B in Figure 5), rivals who have large TPD (Type A in Figure 5) have not made any significant deployments in these technological areas. This absence of rival engagement provides us with a strategic advantage, which can be used as diplomatic tools. These technologies primarily encompass advanced sensors, novel materials, and energy storage and conversion technologies, including microelectromechanical systems, optoelectronic sensors, photovoltaic conversion technologies, solar photovoltaic cells, 3D imaging, and nanomaterials, as determined through the distribution of topic keywords.

Conversely, in the case of topics dominated by competitors—specifically Topics 6, 7, and 9 (Topic A in Figure 5), our country and those friendly nations (Topic B in Figure 5) have few deployments on those topics. If competitors impose export restrictions on these technologies, we may face significant vulnerabilities, potentially leading to supply chain disruptions. These technologies can therefore be identified as high-risk bottleneck technologies with the potential to pose critical challenges to technological and economic security. By reading the keywords identified by TF-IDF algorithm in those topics, it can be found that potential bottleneck technologies are mainly distributed in: **(1) basic electronic components**, including the application of traditional materials such as silicon-based semiconductors and compound semiconductors, **(2) Circuit manufacturing and design**, encompassing ASIC design, chip manufacturing, and packaging technologies, **(3) Signal processing and voltage control**, including analog and digital signal processing technologies used in communications and data processing.

Validation

To validate our findings, we first engaged domain experts in the integrated circuit (IC) industry who hold Ph.D. degrees in semiconductor-related fields and possess both academic and industrial experience. Their combined expertise enables them to make well-informed judgments on the technological landscape. The experts concurred with our conclusions that basic electronic components, circuit manufacturing and design, and signal processing and voltage control constitute China's current bottleneck technologies, primarily controlled by the United States and Japan. These constraints have significantly disrupted China's ability to manufacture advanced chips. However, the experts also noted that due to the vast scope of the IC industry, it is challenging for any single expert to maintain a comprehensive and systematic understanding of the entire technological landscape. As a result, they recommended an additional validation step—comparing our findings with the export control policies of major countries. Following this recommendation, we referenced the U.S. Commercial Control List and its annotation system from the Export Control Database of the National Science Library of Chinese Academy of Sciences (Fang et al., 2022). By analyzing controlled technologies in the integrated circuits sector, we identified the five most highly regulated technologies on the control list: (1) Semiconductor device testing, (2) Electronic testing, (3) Electronic sensors, (4) Communication testing equipment, (5) Wafer inspection-related technologies. All five of these technologies were successfully identified through our methodology. Notably, the electronic sensor technology listed in the control database includes the optoelectronic sensor technology identified in our study. Although subject to export controls, this technology remains an area where China currently holds a competitive advantage, making it less susceptible to becoming a critical bottleneck. In contrast, the other technologies on the control list represent key bottleneck areas that could significantly impact China's technological and industrial security. These results further validate the scientific rigor and practical value of the methodology proposed in this study.

Conclusion and Discussion

In this paper, we propose a novel approach to quantifying the political attributes of technology within the context of global competition. By introducing the concept of political distance, we aim to identify potential bottleneck technologies that may pose risks to national security and highlight technological vulnerabilities. First, we define political distance by considering the countries of patent assignees and conceptualize the Technology Political Distance Indicator as a measure of a country's preference for conducting technology exports. Second, we treat technology as a complex system represented by a citation-based network. Utilizing PageRank and other network-related indicators, we identify critical nodes (patent sets) whose removal could fragment the network and disrupt technological systems, thereby assessing the impact of technology export controls. Third, leveraging both technology political distance and technology control, we categorize technologies into four distinct types and identify potential bottleneck technologies at both the patent and topic levels. Through an empirical study on integrated circuit technologies, our findings indicate

that China holds a leading advantage in cutting-edge applications such as advanced sensors, novel materials, and energy conversion technologies. However, foundational technologies—including basic electronic components, advanced semiconductor materials, and circuit manufacturing and design—are predominantly controlled by countries with which China has distant political relations. Notably, key areas such as logic circuits and electrode integration remain largely underdeveloped domestically. If access to these foundational technologies were restricted, it could severely disrupt China's industrial and supply chains. As such, these fundamental technologies represent critical bottlenecks that China must address. Our results are validated through expert assessments and cross-referenced with the U.S. Commercial Control List, demonstrating the robustness and practical relevance of our proposed method.

Meanwhile, we acknowledge the potential limitations of our research and propose future directions that warrant further investigation. While our study introduces a novel method for quantifying the political attributes of technologies, thereby enhancing the understanding of the nature and implications of bottleneck technologies, it is important to recognize that bottleneck technologies are inherently complex. Their formation is influenced by multiple interrelated factors, including the foundational scientific knowledge, the structure of the technology supply chain, and the positioning of a given technology within the global value chain. These factors interact in intricate ways and collectively shape the emergence of bottleneck technologies. Therefore, we suggest that future research on bottleneck technology theory should focus on developing a rigorous logical framework and modeling approaches to better explain the dynamic mechanisms underlying the formation of bottleneck technologies. From a practical perspective, researchers should also explore strategies for integrating multi-source data to construct a comprehensive and systematic depiction of the technological landscape, enabling more precise identification of critical bottleneck points.

References

Arthur, W. B. (2009). *The nature of technology: what it is and how it evolves*. Simon and Schuster.

Bailey, M. A., Strezhnev, A., & Voeten, E. (2016). Estimating Dynamic State Preferences from United Nations Voting Data. *Journal of Conflict Resolution*, *61*(2), 430-456. https://doi.org/10.1177/0022002715595700

Beltagy, I., Lo, K., & Cohan, A. (2019). *SCIBERT: A Pretrained Language Model for Scientific Text*. Arxiv. Retrieved 2024-6-17 from https://arxiv.org/pdf/1903.10676

Danae, D., Georges, M., & Lefteris, P. (1995). Determining objective weights in multiple criteria problems: The critic method. *Computers & Operation Research*, *22*(7), 763-770.

Davis, C. L., Fuchs, A., & Johnson., K. (2019). State Control and the Effects of Foreign Relations on Bilateral Trade. *Journal of Conflict Resolution*, *63*(2), 405-438. https://doi.org/10.11588/heidok.00017673

Fan, C., Zeng, L., Sun, Y., & Liu, Y. Y. (2020). Finding key players in complex networks through deep reinforcement learning. *Nat Mach Intell*, *2*(6), 317-324. https://doi.org/10.1038/s42256-020-0177-2

Fang, C., Xuezhao, W., Xiwen, L., Yanpeng, W., & Ming, W. (2022). Research on Classification of Scientific Instruments and Technologies in The Commerce Control List of US Export Control. *World Sci-Tech R&D*, *44*(3), 287-298.

Frakes, M. D., & Wasserman, M. F. (2021). Knowledge spillovers, peer effects, and telecommuting: Evidence from the U.S. Patent Office. *Journal of Public Economics*, *198*. https://doi.org/10.1016/j.jpubeco.2021.104425

Guoxiong, Z., Wei, L., Gai, L., & Guanhua, L. (2021). A Research Based on Careful and Thorough Selection of "Neck-jamming"Technologies Using Delphi Method and AHP: a Case Study in Biomedical Field*. *World Sci-Tech R&D*, *43*(3), 331-343.

Haiqiu, Z., Xiaolin, H., Weisi, L., Guilong, L., & Tingyu, W. (2023). Research on the Selection Method of Key and Core Technology Task Based on MD-AHP. *Journal of Intelligence*, *42*(9), 149-154.

Han, X., Zhu, D., Wang, X., Li, J., & Qiao, Y. (2021). Technology Opportunity Analysis: Combining SAO Networks and Link Prediction. *IEEE Transactions on Engineering Management*, *68*(5), 1288-1298. https://doi.org/10.1109/tem.2019.2939175

Jin, C., Zhen, Y., & Zi-qin, Z. (2020). The Solution of"Neck Sticking" Technology During the14th Five- Year Plan Period: Identification Framework, Strategic Change and Breakthrough Path. *Reform*(12), 5-15.

Lee, P.-C. (2021). Investigating the Knowledge Spillover and Externality of Technology Standards Based on Patent Data. *IEEE Transactions on Engineering Management*, *68*(4), 1027-1041. https://doi.org/10.1109/tem.2019.2911636

Luo, Y. (2022). Illusions of techno-nationalism. *J Int Bus Stud*, *53*(3), 550-567. https://doi.org/10.1057/s41267-021-00468-5

S.Burt, R. (1992). *Strutural Holes: The Social Strcuture of Competition*. Havard University Press.

Schmid, S., Lambach, D., Diehl, C., & Reuter, C. (2025). Arms Race or Innovation Race? Geopolitical AI Development. *Geopolitics*, 1-30. https://doi.org/10.1080/14650045.2025.2456019

Sun, H. (2019). US-China tech war: Impacts and prospects. *China Quarterly of International Strategic Studies*, *5*(02), 197-212.

Taichen, W., Minrong, L., & Zhenbiao, C. (2022). An Empirical Study on Influencing Factors of the Value of University Patent Technology Transfer and Transformation: Based on the Patent Transfer and Transformation Data from 11 First-Class Universities. *Library and Information Service*, *66*(9), 103-116. https://doi.org/10.13266/j.issn.0252-3116.2022.09.011

Tongliang, A., Ge, J., & Dazhong, W. (2023). Measurement of Critical Technologies in China's High-tech Manufacturing and the Catch-up Path: Lithium Battery Industry as an Example. *Economic Research Journal*, *58*(01), 192-209.

Vivoda, V. (2023). Friend-shoring and critical minerals: Exploring the role of the Minerals Security Partnership. *Energy Research & Social Science*, *100*. https://doi.org/10.1016/j.erss.2023.103085

Xue, W. (2022). *Study on the Impact of Business Sophistication on National Innovation Output from the Perspective of Knowledge* Huazhong University of Science and Technology]. Wuhan.

Zhiwei, T., Yuxuan, L., & Longpeng, Z. (2021). Identification Method and Breakthrough Path of" Neck-jamming" Technologies under the Background of Sino-US Trade Friction: a Case of the Electronic Information Industry. *Science & Technology Progress and Policy*, *38*(1), 1-9.

https://doi.org/10.51408/issi2025_032

Quantitative Analysis of IITs' Research Growth and Contributions towards Achieving SDGs

Parul Khurana[1], Kiran Sharma[2], Akshat Nagori[3], Manya[4], Mehul Dubey[5]

[1]*parul.khurana@lpu.co.in*
School of Computer Applications, Lovely Professional University, Jalandhar - Delhi G.T. Road, Phagwara, Punjab – 144411 (India)

[2]*kiran.sharma@bmu.edu.in*
School of Engineering & Technology, BML Munjal University, Gurugram, Haryana-122413 (India)
Center for Advanced Data and Computational Science, BML Munjal University, Gurugram, Haryana-122413 (India)

[3]*akshat.nagori.23cse@bmu.edu.in*
School of Engineering & Technology, BML Munjal University, Gurugram, Haryana-122413 (India)

[4]*manya.23cse@bmu.edu.in*
School of Engineering & Technology, BML Munjal University, Gurugram, Haryana-122413 (India)

[5]*mehul.dubey.23cse@bmu.edu.in*
School of Engineering & Technology, BML Munjal University, Gurugram, Haryana-122413 (India)

Abstract

The Indian Institutes of Technology (IITs) are vital to India's research ecosystem, advancing technology and engineering for industrial and societal benefits. This study reviews the research performance of top IITs—Bombay, Delhi, Madras, Kharagpur, and Kanpur based on Scopus indexed publications (1952–2024). Research output has grown exponentially, supported by increased funding and collaborations. IIT-Kanpur excels in research impact, while IIT-Bombay and IIT-Madras are highly productive but show slightly lower per-paper impact. Internationally, IITs collaborate robustly with the USA, Germany, and the UK, alongside Asian nations like Japan
and South Korea, with IIT-Madras leading inter-IIT partnerships. Research priorities align with SDG 3 (Health), SDG 7 (Clean Energy), and SDG 11 (Sustainable Cities). Despite strengths in fields like energy, fluid dynamics, and materials science, challenges persist, including limited collaboration with newer IITs and gaps in emerging fields. Strengthening specialization and partnerships is crucial for addressing global challenges and advancing sustainable development.

Introduction

India, as an emerging global power, has recognized the importance of research and is making concerted efforts to strengthen its academic and scientific ecosystem. Government initiatives such as increased funding for higher education, the establishment of research-centric policies, and global collaborations reflect this commitment (Raaj, 2024).

Research not only advances scientific understanding but also delivers tangible benefits to society, such as sustainable energy solutions, improved healthcare, and technological innovations. In addition, impactful research promotes economic development by creating industries, generating employment and addressing pressing social issues such as poverty, inequality, and climate change (Saini and Chaudhary,

2020). The ability to translate academic insights into real-world applications directly improves the well-being of communities and strengthens the global standing of the nation (Jalal, 2020).

India's premier institutes, the Indian Institutes of Technology (IIT), have been consistently at the forefront of cutting-edge research. These institutions are known for their contributions to diverse fields such as artificial intelligence, clean energy, biotechnology, and advanced manufacturing (Ghosh, 2021). IITs have not only produced groundbreaking research, but have also cultivated an innovation ecosystem that has led to startups, patents, and technology transfer (Nair, Guldiken, Fainshmidt and Pezeshkan, 2015). In Indian education, IITs symbolize excellence in education and research, often being considered centers of intellectual and technological prowess.

Research conducted by IITs creates a ripple effect on innovation, driving comprehensive growth by seamlessly connecting academic exploration with industrial applications. From designing affordable healthcare solutions for underserved communities to pioneering green technologies that address environmental issues, IIT research exemplifies how academia can contribute significantly to society's welfare and national development (Chatterjee and Sahasranamam, 2014). As India aspires to become a global leader in science and technology, the role of IITs in the advancement of impactful research and the promotion of innovation remains critical, contributing not only to the country's economic growth, but also to its social transformation (Cheah, 2016).

Literature Review

Indian Institutes of Technology (IITs) have consistently been recognized as leaders in innovation, research, and academic excellence. The study by Chaurasia and Chavan (2014) provides an insightful evaluation of the productivity of research and the impact of IIT Delhi over a decade. The study concludes that IIT Delhi's research during this period demonstrated significant growth, with increasing contributions to science and technology. Collaboration and interdisciplinary research emerged as key strengths that enhanced the institution's academic reputation, with an emphasis on international partnerships to boost research impact.

The bibliometric study by Awasthi and Sukula (2020) highlighted the key role of IITs in India's scientific growth and global research, emphasizing quality and highly cited and influential publications. The study by Siddaiah, Gupta, Dhawan and Gupta (2016) analyzes the research output and the impact of the citation of eight newly established IITs during 2010–2014. Although relatively new, these IITs demonstrated an increase in research productivity, focusing on engineering, physics, and materials science. Collaborative efforts, especially with established IITs and international partners, greatly improved the impact of citations and the visibility of research. The study highlights the competitive quality of research from these institutions and recommends strengthening research infrastructure and collaborations to further their contributions to the scientific landscape of India.

Pradhan and Sahu (2018) conducted a bibliometric analysis of IIT research publications indexed in Scopus, highlighting trends in productivity, citation impact, and collaborative patterns. The study reveals a significant growth in the publication output, with engineering and technology leading the research domains. International collaborations and high-impact publications underscore the global relevance of IIT research. Ramesh and Pradhan (2017) conducted a scientometric analysis of engineering research at IIT Madras and IIT Bombay from 2006 to 2015. The study also reports a constant growth in research output, higher shares of engineering disciplines, and a high citation impact. Both institutes had wide international collaborations, which gave high relevance and visibility to their research output. The authors underline the strategic partnership and funding as main drivers for sustaining research excellence.

Unique contributions of every individual IIT in research must be studied in depth, while understanding their strengths and weaknesses, their standing in terms of global competitiveness, and their national priority linkages. It also fosters accountability and informs policy decisions related to funding, collaborations, and resource allocation. IIT Bombay has been a leader in database systems and data management research since the 1980s, contributing significantly to information retrieval and data mining (Chakrabarti, Ramakrishnan, Ramamritham, Sarawagi and Sudarshan, 2013). The IIT Bombay Developmental Informatics Lab focuses on leveraging ICT (Information and Communication Technology) to improve access to information in rural India, addressing critical needs such as agriculture and tribal education (Bahuman, Bahuman, Baru, Duttagupta and Ramamritham, 2007). Hasan and Singh (2015) provide a scientometric analysis of the research output of the leading IITs over five years. The research output of the five best performing IITs accounted for 9. 32% of the total Indian research output, with a maximum of 22 27% articles indexed in 2013. Das, Mandal, Rath and Das (2022) have studied that IIT Hyderabad (26%), with an increase in open-access publications, is the top research institute in India for open-access journals. Ghosh (2021) conducted a bibliometric analysis of research productivity in physics, chemistry, and mathematics at IIT Kharagpur from 2001 to 2020. The study highlights significant growth in publication output, with physics leading in productivity and impact of citations. Collaborative research, both national and international, played a key role in enhancing research visibility and impact. The findings underscore the importance of interdisciplinary collaboration and sustained support for research in fundamental sciences. Bhui and Sahu (2018) conducted a bibliometric study of publications by faculty members in the Department of Humanities and Social Sciences of IIT Kharagpur. The analysis highlights the research output trends, with a focus on journal articles and conference papers. The study reveals increasing interdisciplinary research and moderate citation impact, emphasizing the growing contribution of HSS to IIT Kharagpur's academic landscape.

Similarly, other higher education institutes such as IISER, NIT, etc. were also analyzed in terms of their performance. Solanki, Uddin and Singh (2016) evaluated the research competitiveness of IISERs through publications and citations, highlighting their rapid growth and notable contributions in chemistry, physics, and

biology despite their status as relatively young institutions. (Bala and Kumari, 2013) analyze the research performance of the National Institutes of Technology (NITs) of India from 2001 to 2010 using bibliometric methods. The study underscores the consistent growth in research output, with engineering and technology leading the publication domains. Collaborative efforts, both nationally and internationally, have markedly improved the impact of their work.

IITs collectively exhibit significant research contributions in Computer Science, with robust bibliometric indicators such as citation rates and h-index scores, reflecting their global impact (Singh and Singh, 2019; Arif and Badshah, 2015). Krishna and Chandra (2009); Chandra and Krishna (2010) examined the role of IITs in fostering university-industry collaborations and cultivating an entrepreneurial culture, thus improving India's innovation ecosystem (Prathap, 2018) compared the impact of IIT research with leading global institutions in engineering. Boshoff and de Jong (2020) analyzed the social impact of research by presenting it through results, outcomes, and larger societal benefits. They highlight that such research drives progress in public health, education, and economic development while promoting behavioral changes and improving quality of life.

In line with the 17 SDGs of the United Nations, the research output of IITs has contributed substantially to the tackling of critical global challenges, including healthcare, clean energy, sustainable cities and climate action. These institutions play a central role in the advancement of knowledge, the driving force of technological innovation, and the implementation of sustainable solutions to meet the demands of national and global sustainability agendas Priyadarshini and Abhilash (2020); Singh, Kanaujia, Singh et al. (2022).

Research Gap

Although numerous studies have examined the research productivity and impact of IITs, key aspects are still not adequately explored. Previous literature has predominantly focused on overall productivity, citation metrics, and thematic strengths, often overlooking nuanced areas such as the role of interdisciplinary collaborations, the impact of newer IITs, and the alignment of research with emerging global challenges. Furthermore, there is limited exploration of the contribution of IITs to the achievement of sustainable Development Goals (SDGs).

Unlike previous studies that focus on short-term productivity, this study examines seven decades of research output, capturing historical growth trajectories and transformative periods. Our findings also highlight the gaps in collaborative networks, showing limited partnerships between older and newer IITs. By aligning IIT research contributions with SDGs, the study provides information on their role in addressing global challenges along with their international partners. These contributions offer a deeper understanding of IIT research while addressing critical literature gaps, guiding strategies for enhanced collaboration, emerging trends, and social impact.

Research Objectives

- Analyze the growth trajectory of research publications in the five holder IITs (Bombay, Delhi, Madras, Kharagpur and Kanpur) from 1952 to 2024.

- To examine productivity trends and measure citation impacts of publications for each IIT.
- Identify patterns of inter-IIT and international collaborations and their impact on research outcomes.
- Uncover unique research themes and interdisciplinary approaches among the IITs.
- Evaluate the contribution of IIT research to the achievement of the SDGs.

Methodology

Data for the Indian Institutes of Technology (IIT) collaboration analysis were sourced from Scopus, a leading database for peer-reviewed academic literature. The query was constructed using unique Scopus affiliation IDs for IIT-Bombay (IIT-B), IIT-Delhi (IIT-D), IIT-Madras (IIT-M), IIT-Kharagpur (IIT-KGP) and IIT-Kanpur (IIT-K) to specifically capture publications authored by researchers affiliated with these institutions. The search was further refined to include only three types of documents, articles, reviews, and conference papers, as these represent the core academic outputs that reflect substantial research contributions. The query was executed in September 2024, ensuring that the dataset contained the most recent publications available at the time. The resulting dataset included metadata such as publication titles, authors, affiliations, collaboration details, publication years, document types, etc. Figure 1 demonstrates the data downloading and filtering process.

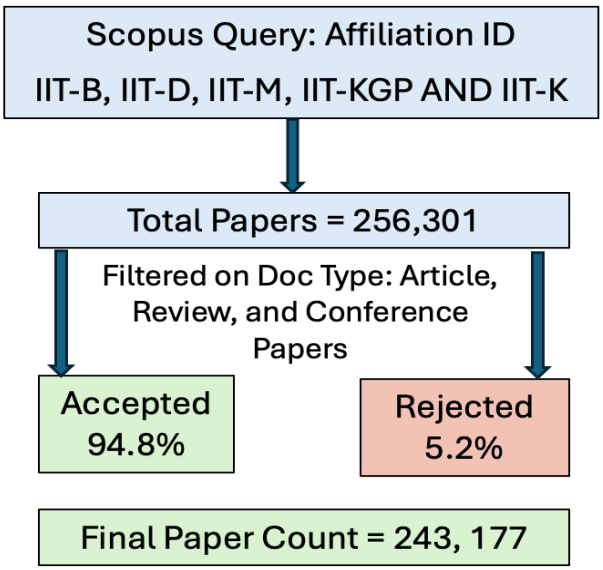


Figure 1. Data downloading and filtering flowchart.

Table 1 provides a data description of research publications from five IITs. Details the year of establishment, affiliation ID, the total number of papers published, the filtered papers, and the percentage of papers accepted and discarded. IIT-KGP

(1951), the oldest institute among the five listed IITs, and IIT-D (1961) is the youngest among five. Publishing in open-access journals often requires article processing charges (APCs), which may not be fully covered by institutional funding, leading to lower open-access publication percentages. Table 2 shows data on the number of research papers published by various IITs and their accessibility. Across all institutes, a smaller percentage of papers (around 10-12%) are open access compared to non-open access, suggesting that most research output is not freely accessible. For IIT-B, open-access papers have a higher citation rate than non open-access papers, suggesting that they might be reaching a broader audience or have higher visibility. In general, non-open-access papers have slightly higher citation rates than open-access papers, possibly because they are published in journals with established academic readerships.

Table 1. Data description.

Institute Name	Abbv.	Estb Year	Affiliation ID	Total Papers	Filtered Papers	% Accepted	% Discarded
Indian Institute of Technology-Bombay	IIT-B	**1958**	60014153	50257	47785	95.08	4.92
Indian Institute of Technology-Delhi	IIT-D	**1961**	60032730	56775	53548	94.32	5.68
Indian Institute of Technology-Madras	IIT-M	**1959**	60025757	52386	50027	95.50	4.50
Indian Institute of Technology-Kharagpur	IIT-KGP	**1951**	60004750	56180	53232	94.75	5.25
Indian Institute of Technology-Kanpur	IIT-K	**1959**	60021988	40703	38585	94.80	5.20

Table 2. Distribution of open access and no open access publications of top 5 IITs.

Institute	Total Papers	Open Access				No Open Access			
		Paper Count	%Paper Count	Total Citations	Paper-Citation Ratio	Paper Count	% Paper	Total Citations	Paper-Citation Ratio
IIT-B	47785	6200	12.98	133807	21.58	41585	87.02	732458	17.61
IIT-D	53548	5898	11.01	115373	19.57	47650	88.99	947281	19.88
IIT-M	50027	5926	11.85	98505	16.63	44101	88.15	756223	17.14
IIT-KGP	53232	5499	10.33	109076	19.83	47733	89.67	973355	20.39
IIT-K	38585	4035	10.46	70731	17.53	34550	89.54	741105	21.45

Table 3 presents data on the distribution of academic papers from five IITs. The data is divided into three categories: Articles, Reviews, and Conference Papers. Articles form the majority of publications in all institutes, contributing approximately 70–77% of the total articles. Reviews constitute a much smaller fraction, around 2–3% of the total papers. Conference Papers make up a significant portion, about 20–27%, depending on the institute. IIT-D leads with 55,348 total published papers, and IIT-

K has the lowest total, at 38,585. IIT-KGP and IIT-M are close in total papers, with IIT-KGP (53,232) slightly ahead of IIT-M (50,027). IIT-B ranks fourth with 47,785 total papers. While all IITs show strong research output, IIT-KGP excels in terms of overall research impact (citations), particularly in articles and reviews. IIT-D and IIT-M show balanced contributions across all categories, while IIT-B demonstrates a strong focus on impactful conference papers. IIT-K, although it has published fewer papers, maintained a steady presence and had potential for growth in impact.

Table 3. Distribution of papers as per document type: article, review, and conference paper.

Institute	Total Papers	Article			Review			Conference Paper		
		Paper Count	%Paper Count	Total Citations	Paper Count	% Paper	Total Citations	Paper Count	% Paper	Total Citations
IIT-B	47785	33474	70.05	718661	1224	2.56	71349	13087	27.39	76255
IIT-D	53548	39316	73.42	874685	1748	3.27	107347	12484	23.31	80622
IIT-M	50027	36904	73.77	735398	1102	2.2	61304	12021	24.03	58026
IIT-KGP	53232	41017	77.05	930932	1452	2.73	89064	10763	20.22	62435
IIT-K	38585	29353	76.07	696738	881	2.28	58597	8351	21.65	56141

In addition, IIT-KGP has the highest overall citation-to-paper ratio (22.04), driven by its strong performance in articles and reviews. The citation-to-paper ratio provides an indicator of the average impact of each paper, measuring how frequently each paper is cited. The citation-to-paper ratio highlights IIT-KGP as the leader in research impact, particularly in articles and reviews, while IIT-D excel in reviews and conference papers.

Results and Discussion

Publication trend analysis

Figure 2 shows the long-term growth trajectory of research outputs from major IITs. This demonstrates the yearly trend in the number of research papers published by five IITs from 1952 to 2024. IIT-KGP led in research output initially, as it was established earlier (in 1951). Other IITs, such as IIT-B and IIT-D, show a slower start in research publications, likely due to being established later. All IITs show steady growth in research output, moving from dozens
to hundreds of papers per year. IIT-KGP maintained a leading position during this period, but other IITs began to close the gap. By the late 1980s and early 1990s, IIT-B and IIT-M had established themselves as competitive research institutions, reaching publication levels similar to IIT-KGP. This period marks a consistent increase across all IITs, reflecting their expanding research focus and resources. There is a notable increase in the number of publications across all IITs, especially from 2000 onward, indicating a surge in research activity. By 2008, the publication output for each IIT reaches more than 1000 papers per year, reflecting enhanced research funding, resources, and collaborative projects. The logarithmic scale emphasizes how each IIT has grown from publishing a handful of papers annually

to publishing thousands, marking a substantial rise in their global research contributions.

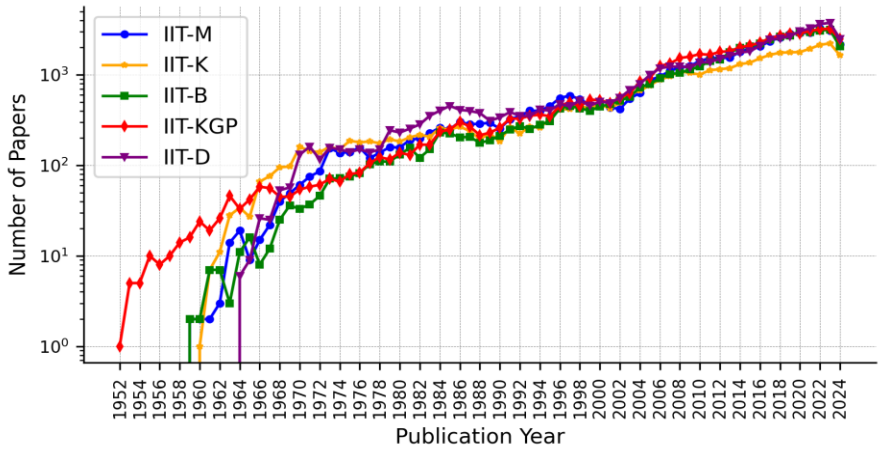


Figure 2. Year-wise publication trends of top 5 ITT's: Madras, Delhi, Bombay, Kharagpur, and Kanpur.

Figure 3 gives a comparative view of research growth in major IITs over the decades. This shows the trend in the percentage of research papers published by five IITs from 1951 to 2024. From 1951 to around 2000, the number of papers published by each IIT remained relatively low and showed only a slight increase. This period reflects the early growth stage of research publications in these institutions. In the 2001–2010 decade, the publication percentages for each IIT began to increase more noticeably, indicating a growth in research output. This decade shows a sharp increase in the number of published papers, with the five IITs reaching their highest publication percentages around 2011–2020. IIT-B, in particular, reached the highest percentage of around 45%, leading the group.

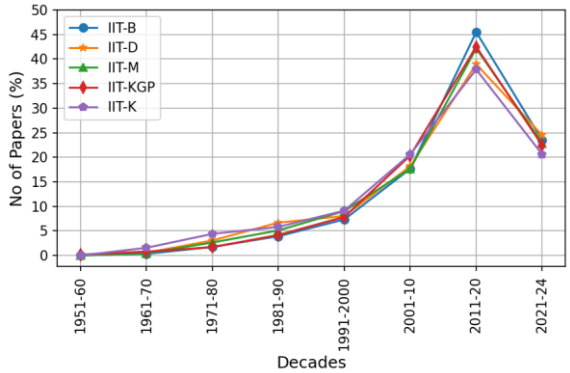


Figure 3. Decade-wise publication trends of top 5 ITT's.

In addition, Table 4 provides a decade-wise distribution of the total number of publications and total citations for the five IITs in the top five from 1951 to 2024.

Each row represents a decade, showing both the number of published papers (Total Papers) and the cumulative number of times, these papers were cited (Total Citations). All IITs show a marked increase in both publications and citations over time, with particularly rapid growth in the 21st century. IIT-B and IIT-M have become leaders in both publications and citations, especially in the last two decades. However, IITs such as IIT-K and IIT-D, despite lower publication counts, have higher citation-to-publication ratios, indicating impactful and highly regarded research. Although IIT-KGP initially led both in publications and citations, IIT-B, IIT-M, and IIT-D have caught up and, in some cases, surpassed IIT-KGP in recent decades, highlighting the dynamic nature of research productivity across IITs. This suggests that, although IIT-K publishes fewer papers compared to some other IITs, its research is highly influential. IIT-KGP follows closely with a ratio of 20.33, showing strong influence and a significant number of citations per paper. IIT-D has a ratio of 19.85, indicating impactful research, slightly lower than IIT-K and IIT-KGP. IIT-B and IIT-M have relatively lower ratios, at 18.13 and 17.09 respectively. Although they have high total citations and publication numbers, their average citations per paper are slightly lower than those of IIT-K, IIT-KGP, and IIT-D. Figure 4 shows the word clouds from five different IITs. Each word cloud represents prominent research themes at each institution, with the size of each word indicating its relative prominence. All IITs share interests in computational methods, optimization, and materials science, each institute has unique specializations that reflect its strengths and research priorities. IIT-B has a balanced focus on materials science (photoluminescence and microstructure) and computational methods (machine learning and optimization), with additional interest in environmental themes like climate change. IIT-D emphasizes energy (Solar Energy, Power Quality), sustainability, and materials science, along with advanced computational methods. IIT-M shows strong interest in fluid dynamics (CFD), materials (microstructure), and computational techniques (Finite Element Method, Optimization). IIT-KGP has a diverse range of themes, but places emphasis on adsorption processes, mechanical properties, and machine learning. IIT-K focuses on corrosion, mechanical properties, and materials science, with a significant presence of computational methods such as finite element analysis. Das (2002) revealed that IIT-Delhi has been actively researching and developing low-emission hydrogen powered engines for nearly two decades, with significant advancements in performance, emission and combustion characteristics.

Table 4. Decade-wise distribution of number of publications and total citations for five IIT's.

Decade	IIT-B		IIT-D		IIT-M		IIT-KGP		IIT-K	
	Total Papers	Total Citations	Total Papers	Total Citations	Total Papers	Total Citations	Total Papers	Total Citations	Total Papers	Total Citations
1951-60	4	126	0	0	4	0	93	915	1	33
1961-70	158	1788	307	1811	234	1509	424	3713	601	10558
1971-80	838	9382	1636	16961	1324	13513	902	8960	1696	31879
1981-90	1872	22801	3580	56129	2551	28480	2226	30658	2244	31866
1991-2000	3509	89725	4309	94486	4528	96939	4177	85433	3524	97981
2001-10	8438	242567	9733	321868	8849	249262	10841	358269	7954	294720
2011-20	21740	438399	20856	477319	21089	403278	22624	508498	14616	299083
2021-24	11226	61477	13127	94080	11448	61747	11945	85985	7949	45356
Total	47785	866265	53548	1062654	50027	854728	53232	1082431	38585	811476



Figure 4. Thematic analysis of research produced by all IITs.

Inter-IIT collaboration

Figure 5 displays the collaboration counts between five IITs - Bombay, Delhi, Madras, Kharagpur and Kanpur and other IITs. Each bar plot represents the number of collaborations between a specific IIT and other IITs, sorted in descending order. Across the five IITs (Bombay, Delhi, Madras, Kharagpur, and Kanpur), collaborations are strongest with each other (the older IITs). IIT Bombay, Kanpur, and Delhi frequently appear as top collaborators in different IITs. IITs Roorkee and Guwahati also have significant collaborations, particularly with Delhi, Madras, and Kharagpur, but tend to have fewer collaborations with newer IITs. IITs like Gandhinagar, Ropar, Jodhpur, and Mandi generally have fewer collaboration counts, indicating fewer interactions with the more established IITs.

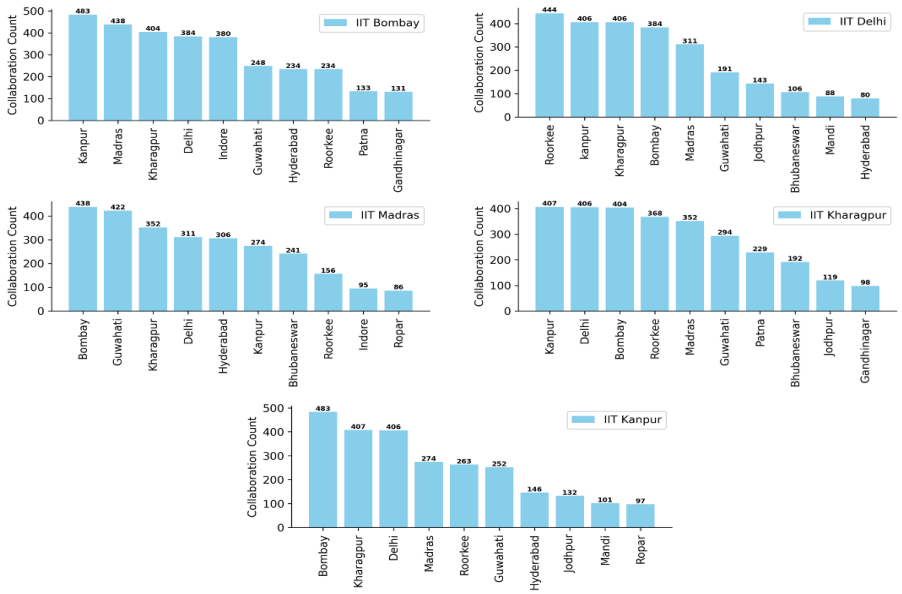


Figure 5. Collaboration of top 5 IITs with other IITs.

IIT-Bombay has its strongest collaborative relationships with IIT Kanpur, Madras, and Delhi, indicating a concentration of joint research or projects with these institutions. IIT-Delhi has its highest collaboration with IIT Roorkee, suggesting strong ties. Its collaborations are relatively well distributed among other major IITs, with lower interaction with newer or smaller IITs. IIT-Madras has strong collaborative links with IIT Bombay, Guwahati, and Kharagpur. There is less collaboration with newer IITs like Ropar and Indore. IIT-Kharagpur collaborates most frequently with IIT Kanpur, Delhi, and Bombay, showing a strong connection with the older IITs. IITs like Gandhinagar and Jodhpur have relatively fewer collaborations with Kharagpur. IIT Kanpur has its highest collaboration with IIT Bombay, which shows a strong link. The lower collaboration counts with newer IITs, like Mandi and Ropar, indicate a preference for working with the established IITs. Overall, this analysis shows a pattern in which older and more established IITs

(Bombay, Delhi, Kanpur, Kharagpur and Madras) tend to collaborate more with each other, probably due to historical relationships, larger research output and available resources. Collaborations with newer IITs are generally lower, which could be due to geographical distance, newer research programs, or differing research focuses.

Authors team size analysis

Teams of two authors make a significant contribution, with IIT-M (35. 18%) and IIT-K (32. 57%) being the highest contributors. IIT-D (29.98%) and IIT-KGP (29.9%) are close, while IIT-B has the smallest percentage for this category (28.5%). The team comprising 3-5 authors is the most dominant category, with percentages exceeding 50% for all IITs. IIT-D leads with 55.51%, followed by IIT-M (50.85%) and IIT-KGP (54.15%). This suggests that most academic papers in these IITs are written collaboratively by mid-sized teams. Extremely large teams (100 authors) contribute a negligible percentage in all IITs, with values not exceeding 1%. IIT-B (0.96%) contributes the most in this category, followed by IIT-M (0.67%). In general, IIT-B appears to have the most diverse team size distribution.

Authorship position distribution

In research, it is often assumed that the first author plays a leading role in the research and writing process (Persson, 2001; Zbar and Frank, 2011). Figure 6 shows the distribution of contributions of authors in different authorship positions (e.g. single author, first author, middle author, last author) in five IITs. The largest contribution across all institutes is made by first authors and middle authors consistently contribute a moderate percentage (around 15% to 17%
across institutes). The contribution of the Last Authors is relatively consistent across institutes, hovering around 9% to 10%. The single authored paper contributed by all IITs is a small proportion with values in the range of approximately 3% to 6%. Overall, the distribution of author positions highlights the hierarchical nature of academic contributions in these institutions, where the first authors play the most significant role.

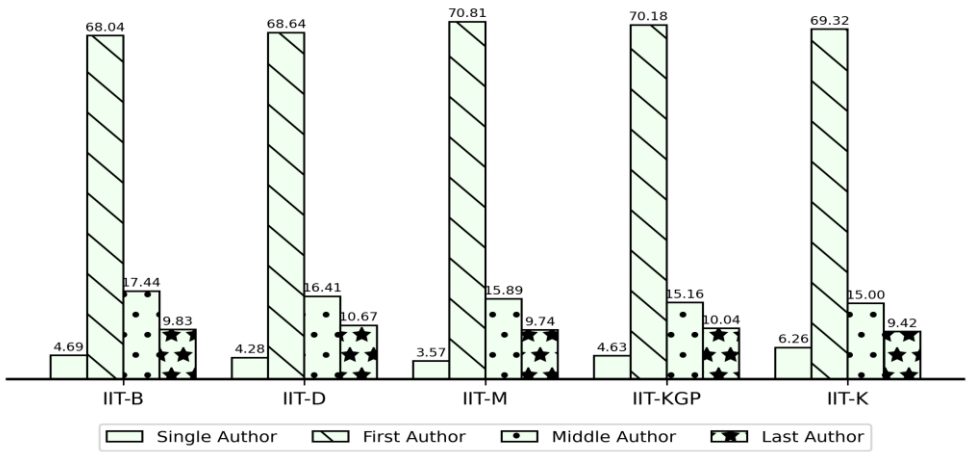


Figure 6. Distribution of papers based on authorship position.

Top authors as per publications

Prof. B.P. Singh from IIT-D has the highest number of publications among all top researchers followed by B. K. Panigrahi from IIT-D, also. Top two authors from IIT-B and IIT-D are from Electrical Engineering, IIT-M and IIT-K are from Chemistry and Electrical Engineering, respectively, and IIT-KGP from Mechanical Engineering and Materials Science & Engineering.

Country collaboration

Figure 7 shows the bar charts of international collaboration counts between five IITs and other top collaborating countries. Each bar graph highlights the number of collaborative research publications between each IIT and different countries, sorted by the most frequent collaborators. The USA is a major collaborator, significantly more than other countries, which suggests that IIT-Bombay has a strong relationship with American institutions. Europe also has a strong presence, especially Germany and the UK. IIT-Delhi also has strong collaborations with the USA, UK, and Germany, although the volume with the USA is less than that of IIT-Bombay. The collaborations are similarly diverse, covering a wide range of countries in Europe, North America, and Asia. IIT-Madras has a strong link with the USA, and Germany, but there is also considerable collaboration with Japan and China, which may indicate research areas where collaboration with Asian countries is important. Unlike other IITs, IIT-Kharagpur has a strong preference for collaboration with the USA. The distribution is slightly less diverse, with fewer collaborations in East Asian countries compared to other IITs like IIT-Madras. IIT-Kanpur shows a similar pattern, with the USA as the primary collaborator, followed by Germany and the UK. East Asian countries have lower collaboration counts, which may indicate a focus on collaborations with North American and European countries.

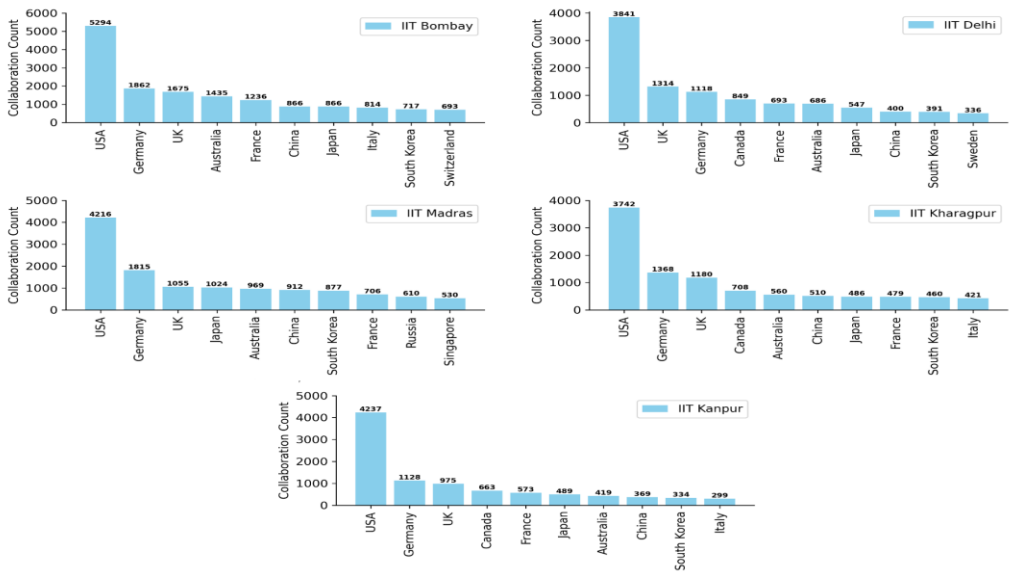


Figure 7. Country collaboration of five IITs.

Overall, the USA is the top collaborator for all IITs, with counts significantly higher than those for other countries, showing a strong reliance on American research institutions and funding sources for collaborative projects. Germany and the UK consistently appear among the top three collaborators for each IIT, suggesting a strong relationship with European institutions. Australia, Japan, and China are important collaborators, particularly for IIT-Bombay, IIT-Madras, and IIT-Delhi, indicating their participation in diverse international research networks. Collaborations with countries such as South Korea, Italy, and Russia vary between IITs, with some showing relatively low counts, especially IIT Kanpur and IIT-Kharagpur. Each IIT shows some variation in its international collaboration patterns, but the overall trend emphasizes North American and European partnerships, with emerging collaborations in Asia. This pattern may reflect differences in research funding, faculty exchanges, and focus areas across IITs.

Proportion of authorship position in top 3 countries

Figure 8 represents a comparison of authorship position top five IITs with their top three collaborating countries: the USA, the UK, and Germany. Each bar is further categorized into three authorship positions: First, Middle, and Last. In the USA, IIT-KGP and IIT-K leads in the first authorship position and IIT-B and IIT-D in second authorship position. The last authorship position consistently has the lowest contribution. In the UK, IIT-B has the highest percentage of the middle category, and IIT-K leads in the first authorship position, while the last category remains minimal, with IIT-B at just 6.99%. In Germany, IIT-KGP dominates the first authorship position and IIT-B leads the second authorship position. Overall, with USA collaboration, first authorship position dominates, with UK and Germany its mix of both first and middle authorship position, and the last authorship position is lowest with all country collaboration.

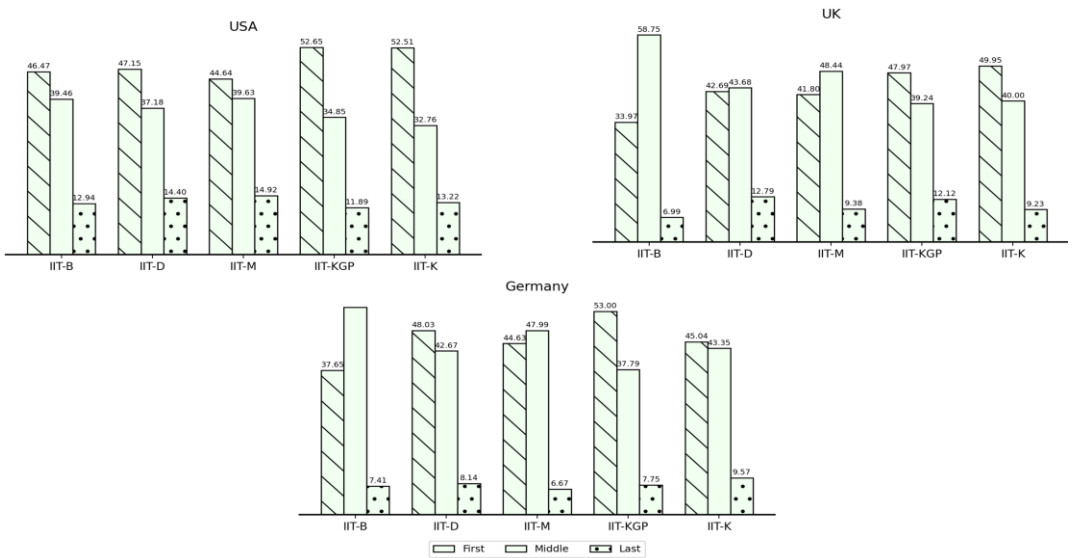


Figure 8. Distribution of authorship position for top three country collaborators.

Collaboration in different regions and continents

Figure 9 shows the distribution of research output across continents that shows consistent patterns among the five IITs, with Europe dominating collaborations. IIT-B shows the highest collaboration with Europe at 48.7%, followed by IIT-M (40. 6%), IIT-K (37. 4%), IIT-D (37%) and IIT-KGP (37%). North America plays a secondary role in collaborations, contributing around 24–30% for most IITs, with IIT-K and IIT-D showing slightly higher engagement. Asia
emerges as the third significant collaborator, with IIT-M (27.1%) and IIT-D (23.8%) exhibiting the highest partnerships in this region. Australia, Africa, and South America collectively account for smaller percentages, typically less than 5% each, reflecting existing but limited partnerships in these regions. IIT-B and IIT-M show slightly stronger collaborations with South America and Australia compared to the other IITs. This comparative analysis highlights Europe and North America as dominant contributors to IITs' research networks while pointing to potential growth opportunities in Asia and emerging collaborations in underrepresented regions like Africa and South America.

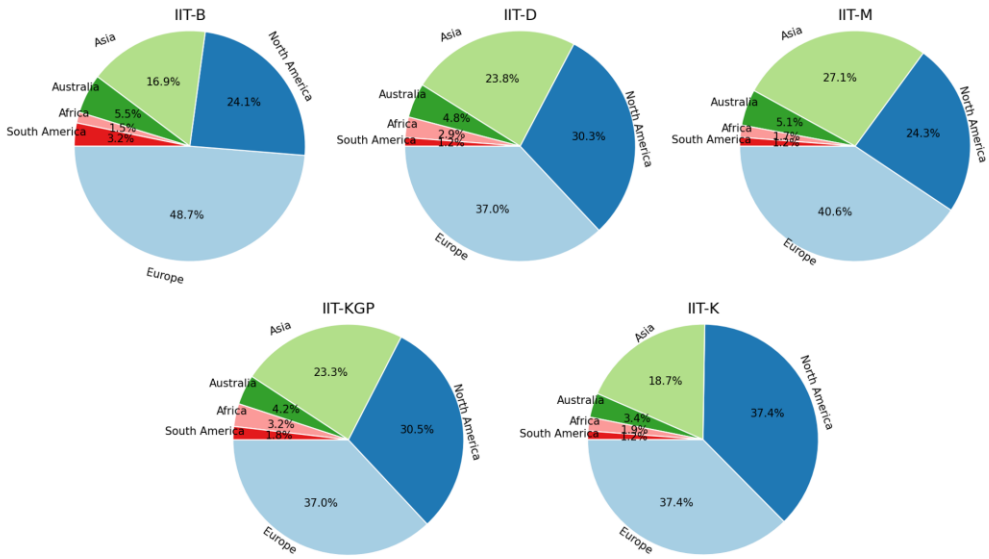


Figure 9. Research output distribution by continent across 5 IITs.

Funding analysis

Funding is essential for academic and scientific growth, especially in research-intensive institutions such as IITs. Funding drives research growth by enabling advanced infrastructure, attracting talent, supporting innovation, fostering collaborations, and improving research quality. It benefits society by addressing critical issues and boosts institutional reputation, creating a cycle of improvement and impact. Table 5 shows results on the research output supported by various prominent funding bodies across five IITs). The table provides the number of papers funded by each organization at each IIT and the total number of citations received

by the papers funded by each organization at each IIT. Each IIT has different levels of support from various funding agencies, reflecting their specific research focuses. DST India and IIT internal funding are the most prominent funding sources, with the highest research output and
impact across all IITs. IIT-KGP shows strong collaboration with NSF, leading to high impact, while other IITs vary in NSF-supported research. Funding bodies like CSIR, SERB, and UGC provide broad and consistent support across IITs, while bodies like MEITY and DRDO have smaller impacts.

Table 5. List of prominent funding bodies in India.

Funding Bodies	IIT-B		IIT-D		IIT-M		IIT-KGP		IIT-K	
	#Papers	TC	#Papers	TC	#Papers	TC	#Papers	TC	#Papers	TC
DST India	2851	43262	3084	49121	2885	47621	2688	50037	2163	45432
IIT	1223	15077	751	12140	1400	15492	1244	21622	1663	22909
CSIR	1080	20475	1259	32985	721	12909	1182	31431	682	14472
NSF	745	45618	394	15319	602	25161	296	9689	370	9792
SERB	735	7196	761	8971	606	5867	761	10233	721	7366
DBT	383	6127	538	11993	294	4270	342	14573	151	3006
DST Kerala	340	7595	388	10366	357	9037	306	10573	368	11258
UGC	324	3598	539	9099	258	3383	144	4128	152	2183
MoE	222	1575	264	1540	247	1105	431	2930	91	498
MHRD	219	4575	225	5827	204	3937	567	14682	141	42042
MNRE	204	4250	69	2668	35	1489	52	1796	41	1058
ISRO	172	1893	31	221	68	867	132	2486	120	2163
DAE	166	2096	73	1032	102	1913	142	3348	125	1988
MEITy	144	672	97	1048	86	422	94	426	104	545
DRDO	107	1371	216	3062	235	4915	187	3704	100	2571

SDGs contribution

Research on SDGs is essential because it addresses global challenges such as poverty, inequality, and climate change by providing evidence-based solutions and driving innovation for sustainability. It informs policies, fosters interdisciplinary collaboration, and helps monitor progress toward these goals. By guiding resource allocation and creating resilience and equality strategies, SDG research ensures a sustainable future for all (Sachs, Schmidt-Traub, Mazzucato, Messner, Nakicenovic and Rockström, 2019; Assembly, 2015). Table 6 provides data on the research output of five Indian Institutes of Technology (IIT), specifically analyzing the extent to which their articles are mapped to the Sustainable Development Goals (SDGs). IIT-D and IIT-KGP perform better in aligning their research output with SDGs compared

to other IITs. IIT-K lags behind in SDG mapping, with the lowest percentage of mapped papers. Across all IITs, the average percentage of papers mapped to SDGs is approximately 35%, indicating a significant gap in research alignment with SDGs. While the IITs are contributing significantly to research, there is substantial room for improvement in mapping their output to SDGs. Encouraging more research initiatives focused on sustainability could enhance alignment with global development goals.

Table 6. SDG count as per each IIT. The values colored in green highlighting the major contributions. Darker the color higher the contribution and vice versa.

IIT	Total Papers	Mapped with SDG			
		Yes		No	
		Count	in %	Count	in %
IIT-B	47783	16781	35.12	31002	64.88
IIT-D	53547	21305	39.79	32242	60.21
IIT-M	50025	16138	32.26	33887	67.74
IIT-KGP	53232	19834	37.26	33398	62.74
IIT-K	38584	10984	28.47	27600	71.53
Total	243171	85042	34.97	158129	65.03

Table 7 provides a detailed breakdown of the contributions of five IITs to the 17 SDGs based on the number and percentage of research papers associated with each goal. IITs collectively focus heavily on SDG 3 (Good Health and Well-being), SDG 7 (Affordable and Clean Energy) and SDG 11 (Sustainable Cities and Communities). IIT-M leads in health-related research with 18. 63% of its articles aligned with SDG 3, while IIT-D excels in energy research, contributing 23. 85% of its articles to SDG 7. Urban sustainability (SDG 11) is a shared focus in all IITs, with contributions averaging around 13%. In contrast, there is minimal focus on SDG 1 (No Poverty), SDG 5 (Gender Equality), and SDG 16 (Peace, Justice and Strong Institutions), with each receiving less than 1% contribution in all IITs. IIT-D shows strength in responsible consumption (SDG 12), IIT-KGP leads in water-related research (SDGs 6 and 14), and IIT-K stands out in food security (SDG 2). Despite significant contributions to key SDGs, IITs have opportunities to expand their focus on social and institutional goals such as eradicating poverty, gender equality, and building peace. The main themes reported by each IIT in SDG 3 and SDG 7 are shown in figure 10.

Table 7. Contribution to SDG goals by five IITs.

SDG Goal	IIT-B		IIT-D		IIT-M		IIT-KGP		IIT-K	
	Count	%	Count	%	Count	%	Count	%	Count	%
1	166	0.64	186	0.55	141	0.59	183	0.56	73	0.45
2	3135	12.05	3339	9.8	2606	10.9	3707	11.27	2122	13.1
3	4601	17.68	5240	15.37	4453	18.63	5327	16.19	2985	18.42
4	589	2.26	463	1.36	441	1.85	587	1.78	303	1.87
5	108	0.42	155	0.45	165	0.69	196	0.6	116	0.72
6	1567	6.02	1763	5.17	1377	5.76	2398	7.29	903	5.57
7	4474	17.19	8127	23.85	4339	18.15	5029	15.29	2824	17.43
8	1197	4.6	1507	4.42	1140	4.77	1481	4.5	770	4.75
9	503	1.93	528	1.55	413	1.73	530	1.61	300	1.85
10	291	1.12	440	1.29	352	1.47	443	1.35	226	1.39
11	3251	12.49	4530	13.29	3137	13.12	4309	13.1	2156	13.31
12	1966	7.56	3391	9.95	1968	8.23	2857	8.68	1252	7.73
13	1300	5	1498	4.4	924	3.87	1568	4.77	554	3.42
14	1167	4.49	1085	3.18	1469	6.15	1870	5.68	784	4.84
15	1338	5.14	1235	3.62	686	2.87	1805	5.49	594	3.67
16	95	0.37	190	0.56	105	0.44	194	0.59	107	0.66
17	272	1.05	405	1.19	185	0.77	412	1.25	132	0.81

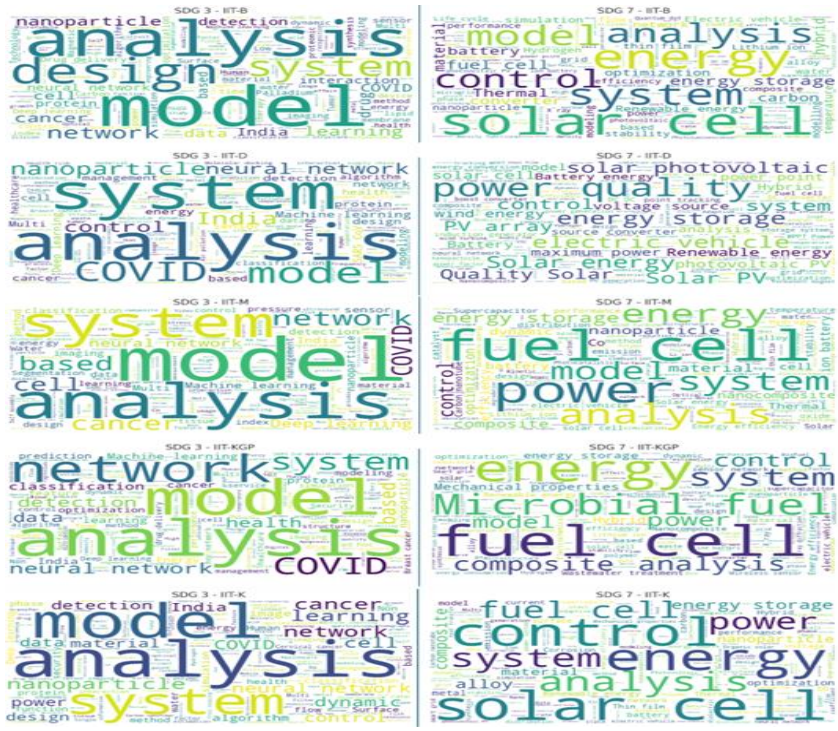


Figure 10. SDG 3 and 7 word cloud.

Country Collaboration: SDG 3 and SDG 7

Figure 11 represents the collaboration counts of five IITs with the top three countries collaborators for two SDGs: SDG 3 (Good Health and Well-being) and SDG 7 (Affordable and Clean Energy). In SDG 3 (Figure 11(left)) shows the highest collaboration counts in all IITs, with IIT-D and IIT-B having particularly strong contributions. UK emerges as the second most significant collaborator for IIT-B, IIT-D, and IIT-K, while Germany is the second most significant collaborator for IIT-M and IIT-KGP. For SDG 7 (Figure 11 (right)), the USA continues its dominance as the main collaborator of all IITs, with IIT-B and IIT-D making substantial contributions. The second and third positions vary by IIT, reflecting differences in regional engagement. This diversity underscores the global nature of SDG-aligned research conducted by the IITs. Across both SDGs, the USA holds the first position for all IITs, underlining its role as the most significant research partner.

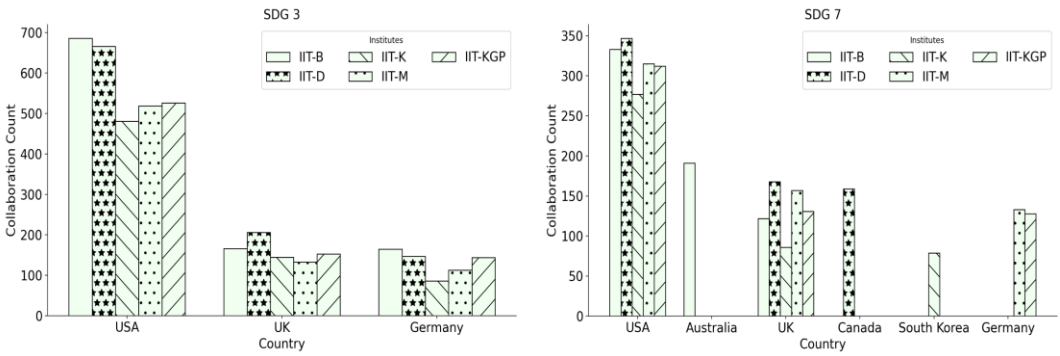


Figure 11. Country Collaboration: SDG 3 and SDG 7.

Conclusion

The Indian Institutes of Technology (IITs) have become the pillars of India's research ecosystem, consistently making significant contributions to global science and technology. Their exponential growth in research productivity over the decades reflects the increasing emphasis on quality, funding, and strategic collaborations. IIT-Kanpur, IIT Kharagpur, and IIT-Delhi have established themselves as leaders in both productivity and impact, while IIT-Bombay and IIT-Madras continue to showcase their dominance in niche areas such as optical materials and fluid dynamics.

Collaborations, particularly with western countries and Asian partners, have significantly increased the global reach and impact of IIT research on the citation. Interdisciplinary approaches integrating computational methods with traditional engineering and sciences highlight the innovative spirit of these institutions. The USA, UK, and Germany are the most frequent collaborators with IITs, underlining their global research connections. Regional collaborators such as Canada, Australia, and South Korea also play significant roles, albeit with specific IITs and research

domains. This reflects the global and goal-specific nature of IITs' international research collaborations.

Collaboration among IITs shows a strong preference for partnerships between older and more established institutions (Bombay, Delhi, Kanpur, Kharagpur, and Madras), likely due to historical ties, increased research capacity, and resource availability. To foster stronger collaboration with newer IITs, older IITs can initiate joint research projects focusing on shared interests such as clean energy, healthcare, and emerging technologies. Resource sharing, including access to advanced laboratories and computational tools, can enable newer IITs to engage in high-impact research. Faculty exchange programs and mentoring initiatives can strengthen the research capabilities of newer IITs, while joint centers of excellence can unite the strengths of both groups to address national and global challenges.

IITs collectively contribute significantly to the global goals of the SDGs, particularly in health, energy, and urban development. However, there is a noticeable lack of focus on social and institutional goals such as poverty eradication, gender equality, and peacebuilding, which presents opportunities for more diversified research in the future.

Limitations

Although the study provides valuable information on research trends and the impact of IITs, certain limitations must be recognized to contextualize its findings. First, the study relies on Scopus-indexed publications, potentially overlooking impactful research published in other indexed databases like WoS, Dimension, and nonindexed or regional journals. Second, only the top five IITs are analyzed, which may not capture the contributions of newer IITs or specific regional variations. Third, the study primarily emphasizes international collaborations and inter-IIT collaboration and may underrepresent domestic or industry-academic partnerships.

Future Scope

- A longitudinal analysis of academia-industry linkages at individual IITs could provide deeper insights into evolving collaboration patterns.
- Applying Social Network Analysis (SNA) could offer a more nuanced understanding of collaboration structures beyond descriptive methods.
- Future research could involve mapping research outputs of IITs to particular SDGs, followed by a comprehensive literature review and adoption of established methodological frameworks used in quantitative SDG-alignment studies.

Acknowledgments

The authors would like to express their sincere gratitude to the reviewers for their valuable feedback and constructive suggestions. The authors also gratefully acknowledge the invaluable learning support provided by the Centre for Teaching, Learning, and Development at BML Munjal University. Special thanks are extended to the Research and Development Cell for their financial support through the seed grant (No: BMU/RDC/SG/2024-06), which made this research possible.

Conflict of interest

All authors declare no conflict of interest.

Data availability

The data utilized in this study is accessible for reproducibility upon request from the corresponding and principal authors.

References

Arif, T., Badshah, B.G.S., 2015. Analyzing research productivity of indian institutes of technology doi:10.5120/cae-1625.

Assembly, G., 2015. Resolution adopted by the general assembly on 11 september 2015. New York: United Nations.

Awasthi, S., Sukula, S.K., 2020. Highly cited publications of selected indian institutes of technology (iits): A bibliometric study. Library Philosophy and Practice, 1–17.

Bahuman, A., Bahuman, C., Baru, M., Duttagupta, S., Ramamritham, K., 2007. Developmental informatics at iit bombay. SIGMOD Record 36, 47–53. doi:10.1145/1276301.1276312.

Bala, A., Kumari, S., 2013. Research performance of national institutes of technology (nits) of india during 2001-2010: A bibliometric analysis. SRELS Journal of Information Management 50, 555–572. doi:10.17821/SRELS/2013/V50I5/43774.

Bhui, T., Sahu, N., 2018. Publications by faculty members of humanities and social science departments of iit kharagpur: A bibliometric study. DESIDOC Journal of Library & Information Technology doi:10.14429/DJLIT.38.6.13569.

Boshoff, N., de Jong, S.P., 2020. Conceptualizing the societal impact of research in terms of elements of logic models: a survey of researchers in sub-saharan africa. Research Evaluation 29, 48–65.

Chakrabarti, S., Ramakrishnan, G., Ramamritham, K., Sarawagi, S., Sudarshan, S., 2013. Data-based research at iit bombay. SIGMOD Record 42, 38–43. doi:10.1145/2481528.2481536.

Chandra, N., Krishna, V., 2010. Academia-industry links: Modes of knowledge transfer at the indian institutes of technology. International Journal of Technology Transfer and Commercialisation 9, 53–76. doi:10.1504/IJTTC.2010.029425.

Chatterjee, D., Sahasranamam, S., 2014. Trends in innovation management research in india - an analysis of publications for the period 1991-2013. Current Science 107, 1800–1805. doi:10.18520/CS/V107/I11/1800-1805.

Chaurasia, N., Chavan, S.B., 2014. Research output of indian institute of technology delhi (iit delhi) during 2001-2010: A bibliometric analysis. International Journal of Information Dissemination and Technology 4, 141–147.

Cheah, S., 2016. Framework for measuring research and innovation impact. Innovation 18, 212 – 232. doi:10.1080/14479338.2016.1219230.

Das, A., Mandal, N., Rath, D.S., Das, S., 2022. Trendline of open access publication by indian institute of technology (iits) researchers in india. SRELS Journal of Information Management doi:10.17821/srels/2022/v59i6/168621.

Das, L., 2002. Hydrogen engine: Research and development (r&d) programmes in indian institute of technology (iit), delhi. International Journal of Hydrogen Energy 27, 953–965. doi:10.1016/S0360-3199(01)00178-1.

Ghosh, T., 2021. Bibliometric investigation on research productivity in physics, chemistry and mathematics in the indian institute of technology (iit) kharagpur during 2001-2020. Indian Journal of Information Sources and Services doi:10.51983/IJISS-2021.11.1.2654.

Hasan, N., Singh, M., 2015. Research output of indian institutes of technology (iits): Ascientometric study. Journal of Knowledge&Communication Management 5, 147–165. doi:10.5958/2277-7946.2015.00012.1.

Jalal, A., 2020. Research productivity in higher education environment. Journal of Higher Education Service Science and Management (JoHESSM) 3.

Krishna, V., Chandra, N., 2009. Knowledge production and knowledge transfer: A study of two indian institutes of technology (iit madras and iit bombay). Information Systems & Economics eJournal doi:10.2139/ssrn.1471105.

Nair, A., Guldiken, O., Fainshmidt, S., Pezeshkan, A., 2015. Innovation in india: A review of past research and future directions. Asia Pacific Journal of Management 32, 925 – 958. doi:10.1007/s10490-015-9442-z.

Persson, O., 2001. All author citations versus first author citations. Scientometrics 50, 339–344.

Pradhan, B., Sahu, S.C., 2018. Bibliometric analysis of research publications of indian institute of technology (iits) based on published literature as reflected in scopus. SRELS Journal of Information Management doi:10.17821/SRELS/2018/V55I5/123101.

Prathap, G., 2018. Comparative evaluation of research in iisc, iits, nus and ntu using cwts leiden ranking 2017 data. Current Science 114, 442–443. doi:10.18520/CS/V114/I03/442-443.

Priyadarshini, P., Abhilash, P.C., 2020. From piecemeal to holistic: Introducing sustainability science in indian universities to attain un-sustainable development goals. Journal of Cleaner Production 247, 119133.

Raaj, S., 2024. Education, research and innovation in india: the shifting paradigms. Journal of Higher Education Theory and Practice 24.

Ramesh, D.B., Pradhan, B., 2017. Scientometrics of engineering research at indian institutes of technology madras and bombay during 2006-2015. DESIDOC Journal of Library & Information Technology 37, 213–220. doi:10.14429/DJLIT.37.3.10967.

Sachs, J.D., Schmidt-Traub, G., Mazzucato, M., Messner, D., Nakicenovic, N., Rockström, J., 2019. Six transformations to achieve the sustainable development goals. Nature sustainability 2, 805–814.

Saini, D., Chaudhary, N.S., 2020. What drives research in higher education? an indian context. Journal of Applied Research in Higher Education 12, 573–584.

Siddaiah, D., Gupta, B., Dhawan, S., Gupta, R., 2016. Contribution and citation impact of eight new iits: A scientometric assessment of their publications during 2010-14. J. Sci. Res. 5, 106–122. doi:10.5530/jscires.5.2.2.

Singh, A., Kanaujia, A., Singh, V.K., et al., 2022. Research on sustainable development goals: How has indian scientific community responded? Journal of Scientific & Industrial Research 81, 1147–1161.

Singh, P.K., Singh, C., 2019. Bibliometric study of indian institutes of technology in computer science, in: 2019 Amity International Conference on Artificial Intelligence (AICAI), pp. 384–393. doi:10.1109/AICAI.2019.8701422.

Solanki, T., Uddin, A., Singh, V., 2016. Research competitiveness of indian institutes of science education and research. Current Science 110, 307–310. doi:10.18520/CS/V110/I3/307-310.

Zbar, A., Frank, E., 2011. Significance of authorship position: an open-ended international assessment. The American journal of the medical sciences 341, 106–109.

Appendix

Table 8. List of all IITs with Scopus affiliation code and abbreviated name.

Code	Abbreviation	IITs
60032730	**IIT D**	Indian Institute of Technology, Delhi
60014153	**IIT B**	Indian Institute of Technology, Bombay
60004750	**IIT KGP**	Indian Institute of Technology, Kharagpur
60021988	**IIT K**	Indian Institute of Technology, Kanpur
60025757	**IIT M**	Indian Institute of Technology, Madras
60031818	**IIT R**	Indian Institute of Technology, Roorkee
60010126	**IIT G**	Indian Institute of Technology, Guwahati
60104350	**IIT I**	Indian Institute of Technology, Indore
60103917	**IIT H**	Indian Institute of Technology, Hyderabad
60104339	**IIT BBS**	Indian Institute of Technology, Bhubaneshwar
60104342	**IIT P**	Indian Institute of Technology, Patna
60104343	**IIT J**	Indian Institute of Technology, Jodhpur
60104341	**IIT GN**	Indian Institute of Technology, Gandhinagar
60104340	**IIT MANDI**	Indian Institute of Technology, Mandi
60103918	**IIT RPR**	Indian Institute of Technology, Ropar
60019106	**IIT V**	Indian Institute of Technology (Banaras Hindu University), Varanasi
60109702	**IIT JU**	Indian Institute of Technology, Jammu
60109689	**IIT PD**	Indian Institute of Technology, Palakkad
60109690	**IIT T**	Indian Institute of Technology, Tirupati
60114558	**IIT GA**	Indian Institute of Technology, Goa
60114557	**IIT BI**	Indian Institute of Technology, Bhilai
60114348	**IIT DHD**	Indian Institute of Technology, Dharwad
60008898	**IIT DBD**	Indian Institute of Technology (Indian School of Mines), Dhanbad

Table 9. List of prominent India funding bodies and abbreviation and full name.

Funding Body	Agency Name
DST India	Department of Science and Technology, Ministry of Science and Technology
IIT	Indian Institute of Technology
CSIR	Council of Scientific and Industrial Research
NSF	National Science Foundation
SERB	Science and Engineering Research Board
DBT	Department of Biotechnology, Ministry of Science and Technology
DST Kerala	Department of Science and Technology, Government of Kerala
UGC	University Grants Commission
MoE	Ministry of Education
MHRD	Ministry of Human Resource Development
MNRE	Ministry of New and Renewable Energy
ISRO	Indian Space Research Organisation
DAE	Department of Atomic Energy, Government of
MEITy	Ministry of Electronics and Information Technology
DRDO	Defence Research and Development Organisation

https://doi.org/10.51408/issi2025_026

R&D Innovation Patterns and Patent Application Strategy of Top-Selling Drugs: Insights from Patentometric

Chao-Chih Hsueh[1], Dar-Zen Chen[2]

[1]*cchsueh@mail.npust.edu.tw*
Department of Business Administration, National Pingtung University of Science and Technology, No. 1, Shuefu Rd., Pingtung, Taiwan (R.O.C)
Center for Research in Econometric Theory and Applications, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C)

[2]*dzchen@ntu.edu.tw*
Department of Mechanical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C)
Center for Research in Econometric Theory and Applications, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C)

Abstract

This article discusses the closed or open innovation patterns adopted by top-selling drugs and their patent application strategy throughout the drug lifecycle. The research samples are 151 top-selling drugs that have ever annual revenue of US $1 billion between 2015 to 2021 identified from the PharmaCompass database and their 1,167 patents listed in the FDA Orange Book. 151 top-selling drugs approved in US FDA from 1988 to 2020. When companies apply a new drug application in the USA, the company needs to submit relevant patents that can reasonably defend against generic drug infringement and list the patent application numbers in the US FDA Orange Book. Besides, we also collected detailed drug lifecycle information from the Orange Book and patentometric information from the USPTO database according to the patent application number listed in the Orange Book. This study uses each new drug's patents listed in the Orange Book, and through the patent holder's information, explores the participant composition in each new drug's R&D process to define the innovation patterns of each new drug. We also compare the innovation patterns' proportions across different drug types. Finally, by utilizing information like the drug's approval date and patent application date, we analyze the differences in patent application scale and patent activity periods across different innovation patterns throughout the drug lifecycle. The results show four innovation patterns. 21.85% of drugs are closed innovation, and the others are open innovation (OI) patterns—30.46% contract, 32.45% coopetition, and 15.23% network open innovation (OI). The top-selling drugs in the general anti-infective disease category have significantly tended to adopt network OI compared to the proportion of antineoplastic and immunomodulating agents and nervous system disease. Besides, from the comparison of patent strategies among the four innovation patterns, the results show that the contract OI drugs have smaller patent scale and shorter patent active period, while on the contrary, network OI drugs have larger patent scale and longer patent active periods. The results provide the pharmaceutical industry with insights into how to use internal and external innovation to find a more efficient and effective R&D management process, diversify a product portfolio to reduce R&D costs, and improve productivity in drug development. Additionally, the study examines the types of patent strategies used to protect drugs under each innovation model.

Introduction

In response to the growing volume and diversity of innovation research in the pharmaceutical industry, Romasanta, van der Sijde, and van Muijlwijk-Koezen (2020) conducted a comprehensive analysis of research topics within innovation

management in the pharmaceutical sector. By employing textual and citation-based clustering analysis on publications from leading innovation management journals, they identified key thematic areas shaping the field. Their findings indicate that strategic alliances have emerged as the most rapidly expanding research focus over the past decade, both in terms of scholarly output and its impact, as measured by citation frequency. Keywords associated with this theme, such as "alliance," "partner," "experience," and "collaborate," underscore the sector's increasing emphasis on cooperative research and development (R&D) initiatives.

The drug discovery and development process is inherently complex, resource-intensive, and time-consuming, requiring a delicate balance between efficacy, safety, regulatory compliance, and commercial viability. Given the substantial financial and operational risks involved, the traditional closed innovation model—where a single firm independently drives pharmaceutical R&D—has increasingly been supplanted by open innovation strategies. This paradigm shift has led to the proliferation of external collaboration mechanisms, including the establishment of dedicated R&D centers, technology licensing agreements, mergers and acquisitions (M&A), and strategic partnerships with competitors and academic institutions (Wellenreuther, Keppler, Mumberg, Ziegelbauer, & Lessl, 2012; Dong & McCarthy, 2019). These collaborative approaches enable firms to leverage complementary expertise, mitigate research risks, and enhance their pharmaceutical product pipelines.

Therefore, many articles within the strategic alliance literature analyze each stage of the alliance, from initiation to management and performance evaluation, while also exploring the factors contributing to its success (Romasanta et al., 2020). However, in the pharmaceutical industry, the product lifecycle of each drug, from R&D exploration, clinical trials, to market launch, can span more than ten years. The information about collaborators or collaboration models through the drug lifecycle may not always be publicly available information. As a result, the research method on pharmaceutical R&D collaboration alliances primarily consists of literature reviews discussing the types of collaborative alliances, case studies examining the management of the collaboration process, or constructing R&D cooperation networks based on publicly available web news. There has been little practical data to verify the innovation performance in the collaboration alliances.

Given that patents offer strong appropriability, we use top-selling drugs with annual revenue of US $1 billion in the U.S. pharmaceutical industry to study the types of R&D innovation patterns of successful drugs based on patentometrics. In the U.S., the Waxman-Hatch Act requires the listing of patents related to each approved New Drug Application in the "Orange Book," including the NDA number, product number, active ingredient(s), trade name, and expiration dates and codes associated with each patent. We can collect patent protection timelines throughout the drug lifecycle of each drug. From a patentometric perspective, we integrate three sources of pharmaceutical data: annual revenue of top-selling drugs (PharmaCompass database), the drug lifecycle information for each drug (FDA Orange Book), and detailed patent information (USPTO). We construct the lifecycles of 151 blockbuster drugs and their patent application timelines and patent applicants. Through a systematic and structural investigation of how pharmaceutical R&D collaboration

works between universities and companies, this research aims to help fill this knowledge gap and provide insights that could enable practitioners to improve the effectiveness of pharmaceutical R&D.

In addition, in the pharmaceutical industry's drug innovation process, different participants possess varying expertise, such as universities engaged in basic research, small companies involved in early drug discovery, and large pharmaceutical companies responsible for late-stage drug development and marketing (Bianchi, Cavaliere, Chiaroni, Frattini, & Chiesa, 2011; Stuart, Ozdemir, & Ding, 2007). Some scholars have explored the driving factors like partner selection from small biotechnology startups or large pharmaceutical companies (Diestre & Rajagopalan, 2012; Mason & Drakeman, 2014). Therefore, this study not only distinguishes the R&D types of new drug development but also explores open innovation R&D models. It examines whether high-profit drug holders get involved in drug R&D process during the research discovery, clinical trial, or market launch phases or if they remain uninvolved in the R&D process and act purely as marketers of the drug. Finally, we also want to study how successful drugs file patents to protect market exclusivity. Different types of R&D innovation patterns exist in the pharmaceutical industry, providing the opportunity to examine how patent protection behavior and time to market differ between closed innovation and open innovation. When drug developers use external knowledge through technology licensing, M&A, and cooperation with competitors or universities, what is the difference in the patent protection strategies among them?

Reference review: R&D innovation patterns in the pharmaceutical industry

Table 1. Comparison of R&D innovation Pattern from reference review.

Author(s)	*Pattern & Definitions*
Felina, E., & Zenger, T.R. (2014)	**Closed Innovation**: Internal innovation processes relying on own resources (e.g. Authority-based, Hierarchy, Consensus-based hierarchy). **Open Innovation**: External innovation processes collaborating with outside parties. (e.g. Markets/Contracts, Partnerships/alliances, Contests/tournaments, Users/communities)
Jackie Hunter and Susie Stephens (2010)	**Closed Innovation:** a model in which firms generate, develop, and commercialize ideas using solely internal resources, maintaining a vertically integrated structure that ensures full control over intellectual property (IP). **Open Innovation:** a paradigm that integrating both internal and external knowledge sources to enhance new product development, foster collaborations with external entities, and enable the commercialization of internal ideas beyond the originating firm.
David Cavalla (2003)	**Contracts:** formalized agreements established to secure external resources necessary for completing specific developmental tasks that cannot be sufficiently addressed internally. These contracts

	emphasize efficient resource allocation and risk mitigation, with compensation typically tied to the completion of designated work, while maintaining minimal dependence on external technology. **Collaborations:** strategic alliances designed to integrate external technologies into an organization's internal discovery processes, thereby enhancing research productivity. **Licensing:** comprehensive agreements that provide access to external products or, in some cases, technologies, to bolster an organization's development pipeline.
Liangsu Wang et al. (2015)	**Traditional Pharma-Academic Partnership:** firms provide financial support to academic researchers in exchange for research outcomes, fostering a structured collaboration aimed at advancing scientific knowledge and achieving specific research objectives. **Open Crowdsourcing:** firms utilize crowdsourcing platforms to seek innovative ideas and solutions from external scientific communities. **Academic Centers of Excellence:** collaborations between pharmaceutical firms and academic institutions, often facilitated by co-located scientists, aim to bridge the gap between academic research and industrial application. **Biotech Co-Creation:** pharmaceutical companies engage with biotech start-ups, pooling resources and expertise to co-develop innovative biotechnological solutions. **Pharmaceutical Peer Risk Sharing:** collaborative ventures between pharmaceutical companies to jointly develop clinical candidates, sharing financial and operational risks in drug development. **Innovation Centers:** pharmaceutical companies establish innovation hubs in key biomedical regions to foster collaborative research, development, and commercialization.
Yeolan Lee et al. (2019)	**Crowdsourcing Open Innovation (OI):** organizations engage in outsourcing problem-solving tasks to leverage collective intelligence to gather novel ideas, solutions, or knowledge, which are then integrated into New Product Development (NPD) processes. **Coopetition Open Innovation (OI):** by sharing resources, expertise, and capabilities across various stages of the NPD or value chain functions, organizations can address complex challenges, overcome limitations, and enhance innovation outcomes. **Science-Based Open Innovation (OI):** companies partner with research institutions such as universities and government laboratories to gain access to cutting-edge scientific knowledge.

	Network Open Innovation (OI): organizations collaborate within networks or consortia to tackle highly complex and interdependent problems. By combining diverse expertise and coordinating efforts across multiple entities.
Alexander Schuhmacher et al. (2022)	**Traditional R&D:** firms primarily rely on internal R&D while selectively incorporating external knowledge through M&A, in-licensing, corporate venture (CV) funds, and collaborations with academia or industry partners. External innovation is limited to portfolio complementation. **Network-Based R&D:** firms expand on traditional R&D by regularly engaging in long-term OI collaborations with multiple partners. **Ecosystem-Enabled R&D:** firms go beyond network-based R&D by leveraging diverse OI processes to acquire technologies and knowledge from multiple sources. They strategically build an open R&D ecosystem, integrating a large number of external contributors.

US FDA orange book and patent linkage system

The Drug Price Competition and Patent Term Restoration Act, commonly referred to as the Hatch-Waxman Act, was enacted in 1984 with the objective of increasing the availability of cost-effective generic drugs to consumers, thereby reducing overall expenditures for U.S. consumers and the healthcare system. Simultaneously, the patent term extension provision within the Act incentivizes brand-name pharmaceutical manufacturers to continue investing in new drug research and development (R&D) by compensating for the regulatory approval timeframe. The Hatch-Waxman Act comprises several key provisions, including the exemption allowing generic drug testing, market exclusivity protections, extensions of patent terms, a streamlined approval process for new drugs, and patent linkage, with the latter being the most intricate and debated aspect. Under the patent linkage framework, the U.S. Food and Drug Administration (FDA) is responsible for compiling and publicly disclosing patent data associated with approved pharmaceutical products, which is recorded in the Approved Drug Products with Therapeutic Equivalence Evaluations, widely recognized as the "Orange Book." When submitting a New Drug Application (NDA) for approval, the applicant must provide not only comprehensive scientific evidence and clinical trial results demonstrating the drug's safety and efficacy but also patent documentation that may serve as a legal basis for preventing generic market entry. This ensures that the FDA includes the listed patents in the Orange Book, allowing for a structured approach to patent enforcement.

Additionally, the Hatch-Waxman Act stipulates that when a generic drug manufacturer submits an Abbreviated New Drug Application (ANDA), it must include one of four specified certifications: (1) Paragraph I, asserting that no relevant patents are recorded in the Orange Book; (2) Paragraph II, indicating that while relevant patents are listed, they have already expired; (3) Paragraph III,

acknowledging the existence of relevant patents but committing to launch the generic drug only after patent expiration; and (4) Paragraph IV, challenging the validity of a listed patent or asserting that the generic drug will not infringe upon it.

The most robust form of patent protection is granted to patents covering the composition of matter, which primarily safeguard the active pharmaceutical ingredient (API) in the drug, followed by patents on novel formulations and drug delivery mechanisms. However, because composition of matter inventions and patent filings for API and original formulations typically occur at the early stages of the drug development cycle, the remaining patent term once the drug reaches the market is often limited, given the extensive time and financial resources required for clinical development and regulatory approval.

The effectiveness of composition of matter patents in protecting repositioned drugs largely depends on whether generic alternatives can be utilized through off-label use to achieve the same therapeutic outcome. In contrast, method of use patents, which cover specific indications or dosing regimens, are often regarded as incremental protections that do not provide the same level of market exclusivity as composition of matter patents. To prolong exclusivity and mitigate the impact of generic competition, pharmaceutical companies continuously invest in R&D throughout the drug lifecycle, securing additional product and method of use patents for the active molecule, thereby reinforcing a comprehensive patent protection strategy.

This study utilizes two primary indicators to evaluate the patent strategies of high-revenue pharmaceuticals: (1) patent scale, denoting the total number of patents registered in the Orange Book, and (2) patent active period, representing the temporal span between the earliest and most recent patent filings within a drug's patent portfolio. However, this analysis does not delve into the specific classifications of patents within each drug's portfolio, such as drug substance patents, product patents, or use patents.

Research Process: identify R&D innovation patterns and their patent protection behavior

This study 151 top-selling drugs with sales of more than one billion US dollars from 2015 to 2021. To search on Drugs@FDA to obtain data such as NDA number, Trade name, Active ingredient, NDA Applicant, IND filing date, NDA approval date, and patent-related information. Then use the Patent number to the USPTO Patent Public Search to search for the Patent applicant and Patent priority date, and integrate the search results into the variables of this study.

Step 1. Collecting patent data

Although the Orange Book offers patent information on each blockbuster drug, it does not contain detailed information on the patents, specifically whether those patents were internally developed by focal organizations or externally sourced. Detailed information on patents is collected from the United States Patent and Trademark Office (USPTO).

Step 2. Identifying R&D innovation patterns

The time and cost risks associated with drugs are very high. Considering the cost risk, in the new drug development process of drugs, in addition to independent research and development, open innovation will also be adopted, such as the establishment of R&D centers, technology licensing, mergers and acquisitions, and cooperation. Innovation and open innovation (OI) to increase the company's pharmaceutical product portfolio. In this study, the research models include closed innovation, contract open innovation, coopetition open innovation, and network open innovation. Open vs. closed innovation choice based on the use of internal or external knowledge in pharmaceutical drug development projects. If drug patents originated from the drug developer they were an internal knowledge source. If drug patents originated from external entities, they were an external knowledge source.

We identified three types of open innovation- contract, coopetition, and network. Biopharmaceutical companies are under immense pressure to improve their R&D productivity. In response, they have increased their portion of outsourced R&D spending on contract research services such as drug discovery, preclinical and clinical activities, or throughout an M&A deal to achieve lower costs, improve speed and flexibility, and minimize risks of new drug development. We called them the Contract OI meaning that the drug developer adopts external knowledge sources completely. Coopetition OI is defined as OI created between firms in the same industry. Coopetition OI can occur between competing firms over different value chain functions or different phases of new product development. Network OI is defined as collaborations between firms and external research organizations including universities, government labs, and other research institutes. External research organizations aim at developing pharma-related knowledge, meanwhile, companies invest in discovering potential scientific collaborators, gaining fundamental scientific knowledge, and turning this into an economic and societal benefit by developing and marketing new drugs (Bekkers & Freitas, 2008; Huang & Chen, 2017). For example, Gleevec was developed between 1987 and 1990 by a team of scientists at Ciba-Geigy in partnership with two researchers at the Dana-Farber Cancer Institute. It is used to treat chronic myelogenous leukemia and was promoted for use by oncologist Brian Druker of Oregon Health & Science University (Druker, 2008; Buchdunger & Zimmerman, 2013). After that, Ciba-Geigy also merged with Sandoz in 1996 to become Novartis, So Gleevec was owned by Novartis and has been registered for a total of 5 patents in the Orange Book. Patent applicants include companies and academic research institutions, such as Novartis, Ciba-Geigy, Dana-Farber Cancer Institute, and Oregon Health & Science University.

Table 2. Definition the closed and open innovation patterns.

		Whether the patent applicant (Drug Patent Holder) is the same as the drug applicant (NDA Applicant)	
		Same (Closed Innovation)	*Different (Open Innovation)*
Patent knowledge Source (patent applicant)	*Internal (I_R&D)*	Closed Innovation	-
	External (E_R&D)	-	Contract Open Innovation Coopetition Open Innovation Network Open Innovation

Step 3. Comparison of therapeutic market classes and patent protection behavior among the different R&D innovation patterns

The patent protection behavior contains four indicators and their definition are as follows. Patent scales are the total number of patents throughout the drug lifecycle; No. of patents before /after NDA is the total number of patents the application date before or after the new drug approved marketing date; the Patent active period is the years between the latest filing date and earliest filing date for patent application. Therapeutic market classes include ten therapeutic market areas-alimentary tract and metabolism (A), blood and blood-forming organs (B), cardiovascular system (C), genito-urinary system and sex hormones (G), systemic hormonal preparations, excluding sex hormones and insulins (H), general anti-infectives for systemic use (J), antineoplastic and immunomodulating agents (L), nervous system (N), respiratory system (R) and others.

The trends of top-selling drugs adopted innovation patterns

Our analysis of the 151 top-selling drugs in the sample revealed that 118 drugs (78.14%) were developed through open innovation (OI) projects, while 33 drugs (21.86%) were developed using closed innovation approaches. Within the OI projects, 30.46% of the drugs were associated with contract open innovation, 32.45% with coopetition, and 15.23% with network open innovation. The data indicates that collaborations between pharmaceutical companies or between pharmaceutical and biotechnology companies are the most prevalent forms of OI models in this sector, with coopetition OI being the dominant model.

OI may take the shape of networks, ecosystems, or consortia, where multiple entities contribute to new product development (Vanhaverbeke & Cloodt, 2006). According to Nambisan & Sawhney (2011), network OI is distinguished by the coordination processes required among multiple organizations within the network, which are necessary to manage the increasing complexity of technological advancements (Vanhaverbeke & Cloodt, 2006). Furthermore, Ritter & Gemunden (2003) assert that network OI is particularly effective in addressing challenges associated with the

intricacies of interconnected technologies. Consequently, network OI represents a relatively smaller proportion of the OI models within the pharmaceutical industry.

Table 3. Definition and example of four R&D innovation patterns.

Research Models (RMs)	*%*	*Drug name*	*NDA Applicant*	*Total patents*	*Innovation Source*	*Drug Patent Holder(s)*[*]
Closed Innovation (1)	21.85	Imbruvica	Pharmacyclics	40	I_R&D	Pharmacyclics
		Jakafi	Incyte	9	I_R&D	Incyte
		Kalydeco	Vertex	11	I_R&D	Vertex
Contract Open Innovation (2)	30.46	Linzess	Allergan	12	E_R&D	Ironwood; Microbia; Ironwood, Forest Laboratories
		Myrbetriq	Apgdi	10	E_R&D	Astellas; Yamanouchi
		Vyvanse	Takeda	18	E_R&D	New River; Shire
Coopetition Open Innovation (3)	32.45	Farxiga	AstraZeneca	37	I_R&D	AstraZeneca
					E_R&D	Alkermes; Amylin; Alkermes, Amylin; Amylin, AstraZeneca; Bristol Myers Squibb; Mitsubishi Electric; TecPharma Licensing
Network Open Innovation (4)	15.23	Genvoya	Gilead	16	I_R&D	Gilead
					E_R&D	Emory University; Japan Tobacco; Brother Kogyo Kabushiki Kaisha
		Gleevec	Novartis	5	I_R&D	Novartis
					E_R&D	Ciba-Geigy; Novartis, Dana-Farber Cancer Institute, Oregon Health & Science University

[*] Comma indicates that there are multiple applicants for the same patent; Semicolon indicates that the patent is from different applicants.

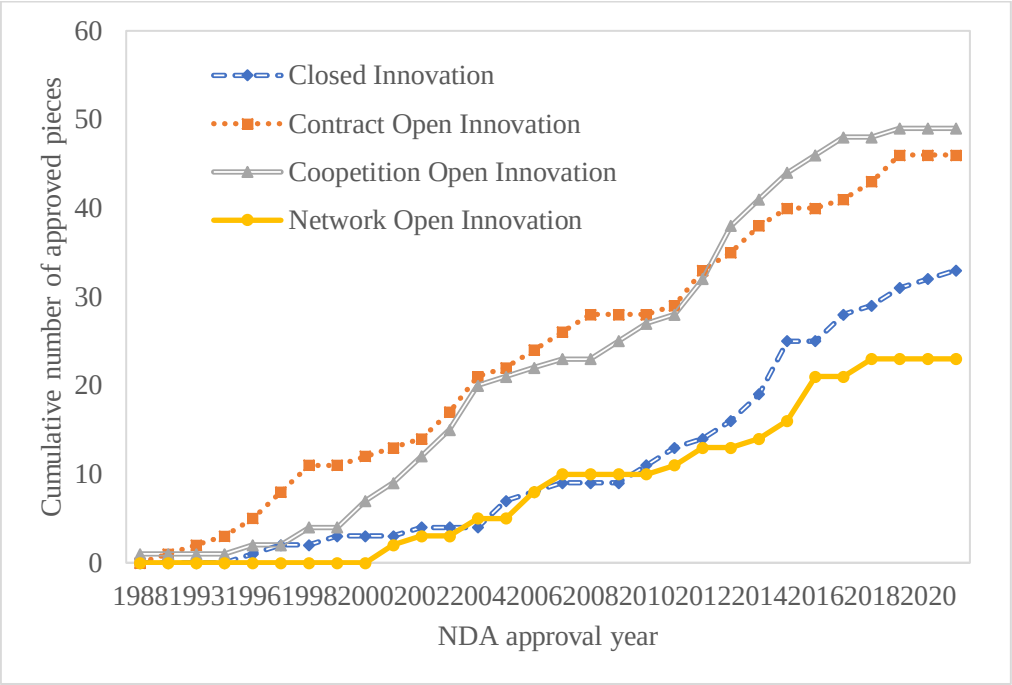


Figure 1. NDA approval cumulative trends of four innovation patterns.

In Table 4 the examination of research and development (R&D) innovation patterns across various Anatomical Therapeutic Chemical (ATC) classifications reveals that the first three categories—Closed Innovation, Contract Open Innovation, and Coopetition Open Innovation—demonstrate relatively consistent adoption rates across different ATC classifications. This uniformity is corroborated by statistical analysis, which shows no significant differences, suggesting that the disease types associated with new drug development within these three categories do not exhibit notable variation in their adoption of R&D innovation patterns.

Specifically, the association between the percentage distribution of Anatomical Therapeutic Chemical (ATC) codes and the network open-innovation model was found to be statistically significant, $\chi^2(27, N = 151) = 56.342$, $p = .001$. Subsequent comparisons of the proportions of ATC codes at the network open-innovation stage, utilizing z-tests, revealed that the proportion of General anti-infectives for systemic use (ATC code J) was significantly higher than that of Antineoplastic and immunomodulating agents (ATC code L) (46.7% vs. 9.8%, respectively) and Nervous system (ATC code N) (46.7% vs. 0.0%, respectively) at the $p < .05$ level.

Table 4. Comparison of drug R&D innovation patterns among therapeutic classes (TC).

TC codes / *R&D models*		*A*	*B*	*C*	*G*	*H*	*J*	*L*	*N*	*R*	*others*	*Total*	*Post-hoc*
(1)	Count	5_a	4_a	1_a	0_a	1_a	1_a	12_a	4_a	4_a	1_a	33	-
	% of RMs	15.2	12.1	3.0	0.0	3.0	3.0	36.4	12.1	12.1	3.0	100	
	% of TC	38.5	44.4	9.1	0.0	25.0	3.3	29.3	21.1	36.4	16.7	21.9	
	% of total	3.3	2.6	0.7	0.0	0.7	0.7	7.9	2.6	2.6	0.7	21.9	
(2)	Count	1_a	3_a	5_a	5_a	1_a	6_a	13_a	8_a	2_a	2_a	46	-
	% of RMs	2.2	6.5	10.9	10.9	2.2	13.0	28.3	17.4	4.3	4.3	100	
	% of TCs	7.7	33.3	45.5	71.4	25.0	20.0	31.7	42.1	18.2	33.3	30.5	
	% of total	0.7	2.0	3.3	3.3	0.7	4.0	8.6	5.3	1.3	1.3	30.5	
(3)	Count	4_a	2_a	5_a	2_a	2_a	9_a	12_a	7_a	5_a	1_a	49	-
	% of RMs	8.2	4.1	10.2	4.1	4.1	18.4	24.5	14.3	10.2	2.0	100	
	% of TCs	30.8	22.2	45.5	28.6	50.0	30.0	29.3	36.8	45.5	16.7	32.5	
	% of total	2.6	1.3	3.3	1.3	1.3	6.0	7.9	4.6	3.3	0.7	32.5	
(4)	Count	$3_{a, b}$	$0_{a, b}$	$0_{a, b}$	$0_{a, b}$	$0_{a, b}$	**14_b**	**4_a**	**0_a**	$0_{a, b}$	$2_{a, b}$	23	J>L,N
	% of RMs	13.0	0.0	0.0	0.0	0.0	**60.9**	**17.4**	**0.0**	0.0	8.7	100	
	% of TCs	23.1	0.0	0.0	0.0	0.0	**46.7**	**9.8**	**0.0**	0.0	33.3	15.2	
	% of total	2.0	0.0	0.0	0.0	0.0	9.3	2.6	0.0	0.0	1.3	15.2	
Total	Count	13	9	11	7	4	30	41	19	11	6	151	
	% of RMs	8.6	6.0	7.3	4.6	2.6	19.9	27.2	12.6	7.3	4.0	100	
	% of TCs	100	100	100	100	100	100	100	100	100	100	100	
	% of total	8.6	6.0	7.3	4.6	2.6	19.9	27.2	12.6	7.3	4.0	100	

Note: A: Alimentary tract and metabolism; B: Blood and blood forming organs; C: Cardiovascular system; G: Genito-urinary system and sex hormones; H: Systemic hormonal preparations, excluding sex hormones and insulins; J: General anti-infectives for systemic use; L: Antineoplastic and immunomodulating agents; N: Nervous system; R: Respiratory system.

Comparison of patent application strategy among the different innovation patterns

In Table 5 and 6, the coopetition and network open innovation (OI) models demonstrate the highest number of patents and the longest patent active periods throughout the drug development lifecycle, surpassing the contract open innovation (OI) model. While the contract OI model exhibits the fewest patents, it is characterized by the shortest R&D time required to bring a drug to market. This study reveals that the four R&D innovation patterns possess distinct characteristics, providing pharmaceutical companies with a range of strategic options to develop their product portfolios. Patents serve a critical role in governing the interactions between various stakeholders in open innovation, particularly by defining and safeguarding technological innovations, such as when large firms acquire startups. In the pharmaceutical industry, where R&D investments are substantial and development timelines are extended, patents are crucial for recouping R&D expenditures, leading to a high propensity for patenting (Arundel, 2001). Small, technology-based firms, often constrained by limited financial resources (Storey & Tether, 1998), tend to prioritize patent filings at later stages, if at all, to minimize costs, which contributes to the smaller patent scale and shorter patent active period observed in the contract OI model.

Table 5. Profile of patent application strategy among four innovation patterns based on ANOVA analysis.

RMs	*No. of patents (N=151)*							
	N	*Mean*	*SD*	*F-value*	*Significance or difference (Dunnett T3-test, p-values)*			*Post-hoc*
					(2)	*(3)*	*(4)*	
(1)	33	11.00	9.57	9.370/5.417 [a]	0.018	1.000	1.000	1,3,4>2
(2)	46	5.61	4.48	(0.000/0.002)		0.007	0.001	
(3)	49	10.41	8.81				0.998	
(4)	23	11.13	5.19					

[a] Welch/Brown-Forsythe, asymptotically F distributed. ** $p<0.05$; *** $p<0.001$.

Table 6. Profile of patent application strategy among four innovation patterns based on ANOVA analysis.

RMs	*Patent active period (N=151)*							
	N	*Mean*	*SD*	*F-value*	*Significance or difference (Scheffé-test, p-values)*			*Post-hoc*
					(2)	*(3)*	*(4)*	
(1)	33	13.35	5.10	12.147	0.108	0.228	0.074	3,4>2
(2)	46	9.70	6.37	(0.000)		0.000	0.000	
(3)	49	16.38	7.24				0.805	
(4)	23	17.99	6.42					

** $p<0.05$; *** $p<0.001$.

Conclusion

The rising pharmaceutical costs and sharply declining R&D productivity have prompted the pharmaceutical industry to seek external innovation models in the hope of producing breakthroughs in the R&D process to reduce R&D costs and improve productivity in drug development. The results in this study provide a reference for pharmaceutical companies to adopt these R&D innovation models, a comparison of patent scale at different life cycle stages, and patent active period among them to get more efficient and effective R&D management process and diversify a product portfolio in drug development.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 113L900202) which is under the Featured Areas Research Center Program by Higher Education Sprout Project of Ministry of Education (MOE) in Taiwan.

References

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research policy*, *30*(4), 611-624.

Bekkers, R., & Freitas, I. M. B. (2008). Analysing knowledge transfer channels between universities and industry: To what degree do sectors also matter?. *Research policy*, *37*(10), 1837-1853.

Bianchi, M., Cavaliere, A., Chiaroni, D., Frattini, F., & Chiesa, V. (2011). Organisational modes for Open Innovation in the bio-pharmaceutical industry: An exploratory analysis. *Technovation, 31(1),* 22–33.

Buchdunger, E., & Zimmerman, J. (2013). The Story of Gleevec. In Innovation.org. Retrieved from (https://web.archive.org/web/20131021011042/http://www.innovation.org/index.cfm/StoriesofInnovation/InnovatorStories/The_Story_of_Gleevec).

Cavalla, D. (2003). The extended pharmaceutical enterprise. *Drug Discovery Today*, 8(6), 267-274.

Diestre, L., & Rajagopalan, N. (2012). Are all 'sharks' dangerous? new biotechnology ventures and partner selection in R&D alliances. *Strategic Management Journal*, 33(10), 1115–1134.

Dong, J. Q., & McCarthy, K. J. (2019). When more isn't merrier: pharmaceutical alliance networks and breakthrough innovation. *Drug discovery today*, *24*(3), 673-677.

Druker, B. J. (2008). Translation of the Philadelphia chromosome into therapy for CML. *Blood, The Journal of the American Society of Hematology*, *112*(13), 4808-4817.

Felin, T., & Zenger, T. R. (2014). Closed or open innovation? Problem solving and the governance choice. *Research policy*, 43(5), 914-925.

Huang, M. H., & Chen, D. Z. (2017). How can academic innovation performance in university–industry collaboration be improved?. *Technological Forecasting and Social Change*, *123*, 210-215.

Hunter, J., & Stephens, S. (2010). Is open innovation the way forward for big pharma?. *Nature Reviews Drug Discovery*, 9(2), 87-88.

Lee, Y., Fong, E., Barney, J. B., & Hawk, A. (2019). Why do experts solve complex problems using open innovation? Evidence from the US pharmaceutical industry. *California Management Review*, 62(1), 144-166.

Mason, R., & Drakeman, D. L. (2014). Comment on "fishing for sharks: Partner selection in biopharmaceutical R&D alliances" by diestre and rajagopalan. *Strategic Management Journal*, 35(10), 1564–1565.

Nambisan, S., & Sawhney, M. (2011). Orchestration processes in network-centric innovation: Evidence from the field. *Academy of management perspectives*, *25*(3), 40-57.

Ritter, T., & Gemünden, H. G. (2003). Network competence: Its impact on innovation success and its antecedents. *Journal of business research*, *56*(9), 745-755.

Romasanta, A. K. S., van der Sijde, P., & van Muijlwijk-Koezen, J. (2020). Innovation in pharmaceutical R&D: Mapping the research landscape. *Scientometrics*, 125, 1801-1832.

Schuhmacher, A., Gassmann, O., Bieniok, D., Hinder, M., & Hartl, D. (2022). Open innovation: A paradigm shift in pharma R&D?. *Drug Discovery Today*, 27(9), 2395-2405.

Storey, D. J., & Tether, B. S. (1998). New technology-based firms in the European Union: an introduction. *Research policy*, *26*(9), 933-946.

Stuart, T. E., Ozdemir, S. Z., & Ding, W. W. (2007). Vertical alliance networks: The case of university–biotechnology–pharmaceutical alliance chains. *Research Policy*, 36(4), 477–498.

van de Vrande, V. (2013). Balancing your technology-sourcing portfolio: How sourcing mode diversity enhances innovative performance. *Strategic Management Journal*, 34(5), 610–621. https://doi.org/10.1002/smj.2031.

Vanhaverbeke, W., & Cloodt, M. (2006). Open innovation in value networks. *Open innovation: Researching a new paradigm*, *13*, 258-281.

Wang, L., Plump, A., & Ringel, M. (2015). Racing to define pharmaceutical R&D external innovation models. *Drug discovery today*, 20(3), 361-370.

Wellenreuther, R., Keppler, D., Mumberg, D., Ziegelbauer, K., & Lessl, M. (2012). Promoting drug discovery by collaborative innovation: a novel risk-and reward-sharing partnership between the German Cancer Research Center and Bayer HealthCare. *Drug discovery today*, *17*(21-22), 1242-1248.

https://doi.org/10.51408/issi2025_079

Rapid Growth of Research Output Amidst Political Instability: A Study of Libya's Last 20 Years

Stephen Wu[1], Adel Diyaf[2], Reem Abusanina[3]

[1]*wu.stephen.t@gmail.com*
Saraya Hamra University, Center for Research & Innovation, Tripoli (Libya)

[2]*a.diyaf@uot.edu.ly*
University of Tripoli, Dept of Biomedical Engineering, Tripoli (Libya)
Saraya Hamra University, International Cooperation Office, Tripoli (Libya)

[3]*r.abusanina@shu.edu.ly*
Saraya Hamra University, Center for Research & Innovation, Tripoli (Libya)

Abstract

Political stability is widely seen as a foundational building block of a national innovation system. For countries like Libya whose last 20 years have included revolution and civil war, this means that low research productivity has been assumed and thus ignored. Seeking to empower the innovation system in Libya, this article examines the current status of Libya's research accomplishments and capabilities through the lens of top-tier scientific research output. Counterintuitively, this retrospective bibliometric study on the Web of Science shows robust research growth in Libya over the last 20 years, even through political turmoil and despite lack of funding. International partnerships are noted as a key correlate of this growth, perhaps supported by capacity building projects and mobility programs. While the overall scientific output from Libya is currently low relative to regional, economic, and developmental comparisons, the growth also suggests the existence of substantial intellectual capital that could sustain expansion in research and innovation.

Introduction

It is widely accepted that key components of national innovation systems (NIS; Lundvall et al., 2002) have difficulty thriving in the midst of political instability (Feng, 1997; Globerman & Shapiro, 2003; Leydesdorff & Meyer, 2006). Libya is widely viewed as such a context, given a societal revolution in 2011 and an unresolved civil war beginning in 2014. Consequently, it has often been left out of studies on scientific productivity, even within its own geographic region (Aggarwal et al., 2020; Ali & Elbadawy, 2021; Landini et al., 2015; Medina, 2015; Radwan, 2018), and also in global indicators such as the World Intellectual Property Organization's Global Innovation Index (GII).

A key part of an NIS is scientific research, where Arab nations face regionally common barriers such as lack of resources, funding, and research infrastructure (Elgamri et al., 2024). These findings were corroborated for the case of Libya in the reports of the recent IBTIKAR project (UNIMED, 2024) – a capacity-building effort, funded by the European Commission under the Erasmus+ program. In surveys, site visits, and training for the 11 participating Libyan universities, Libyan researchers expressed many difficulties in the national and institutional research climate. As part of its efforts to address these issues, the IBTIKAR project provided research

equipment to support and enhance the capabilities of these institutions, aiming to foster a more conducive research climate. The project characterized Libyan research & innovation (R&I) as "embryonic"; they called for (and sought to lead the way to) "a more mature phase" of R&I in Libya. SCIMAGO, based on Scopus-indexed articles, currently ranks Libya as #113 in its Country Rank – second-to-last in the Middle East & North African (MENA) region.

With an overarching goal of empowering the innovation system in Libya, we start by trying to understand the current status of Libya's research accomplishments and capabilities. Our task in this study is to quantify the state of top-tier scientific research output in Libya. In particular, we consider Web of Science (WoS) publications over the last 20 years, and ask:

- RQ1: How does Libya's output in research publications compare with regional and global output?
- RQ2: How has Libya's output in research publications been affected by the sociopolitical environment and events?
- RQ3: How has Libya's output in research publications been affected by (a) funding practices and (b) international partnerships?

With the results of IBTIKAR and anecdotal evidence of the challenges experienced by researchers, we hypothesized that Libyan scientific output would be relatively sparse compared to similar nations, impaired by political instability, poorly funded, and weakly partnered. The results of our study do in fact show that research output is lower than comparison countries along several intuitive axes, and that Libya exhibits low degrees of overall funding for research projects.

However, the major result of this present work is that, counterintuitively, *growth* in top-tier scientific output from Libya in 2004-2024 has far outstripped global comparisons, and has kept pace with North African counterparts. Furthermore, the longstanding political instability and crisis-level national events such as revolution and war have only minor, short-term effects. This growth persists despite few increases in domestic funding, in that most publications do not report any funding sources.

Another result of our study is that partnerships with international entities have been very important for Libyan research. This is especially reflected in the composition of authorial teams, the imperviousness of funded projects to political turmoil, and the decreasing indigeneity of Libyan research. We conjecture about how indirect funding, through capacity-building and mobility programs, may contribute to the creation of an alternative structure within the Libyan NIS.

Finally, we suggest that Libya's research capabilities are strong and severely under-utilized within the existing available NIS. We also suggest some follow up work in Libyan innovation studies.

Data and Methods

We first sought to compare Libyan WoS publications to regional, economic, developmental, and population-size counterparts (RQ1, results in Table 1). Then we considered how Libyan research output has developed over time, compared to aggregated comparison countries (RQ1 & RQ2, results in Figure 1); how domestic

vs. international funding has correlated with WoS output (RQ3, results in Table 2 and Figure 2); and how domestic vs. international teams have correlated with WoS output (RQ3, results in Table 3 and Figure 3).

Inclusion Criteria

To form the body of scientific literature for analysis, we accessed the Web of Science (WoS) on November 14, 2024 and searched for "Address" to include "Libya"; the "Year Published" to range from 2004-2024, and the "Document type" to be articles, proceedings papers, book chapters, or review articles. With this time range, we retrieved a resulting 7,821 WoS-indexed articles.

Aside from Libya, we compared with other countries or groups of countries. In doing so, we followed the same procedure on December 4, 2024 as we did for Libya, except that we listed those comparison countries under "Address" (and in the case of global-scale WoS statistics, we removed the "Address" requirement).

Factors for heuristic comparison

We performed comparisons of Libya with other countries and regions, utilizing heuristic factors (i.e., common-sense labels) that were defined as follows:

- **Regional**. North Africa: Mauritania, Morocco, Algeria, Tunisia, Libya, and Egypt.
- **Economic**. Gross Domestic Product (GDP): From the World Bank 2023.[1]
- **Development**. Human Development Index (HDI): From the United Nations Development Programme (UNDP) 2022.[2]
- **Population**: From the United Nations Department of Economic and Social Affairs (UN DESA) 2023.[3]

For each of the quantitative factors above, we considered the global pool of ranked countries, and chose the two countries that were numerically above Libya and the two countries that were numerically below Libya. For example, the Economic comparison group consisted of 4 countries: 2 with GDP just above Libya (Turkmenistan and Jordan), and 2 with GDP just below Libya (Uganda and Tunisia). In addition to comparing Libya vs. the other countries, this elucidates which factors are salient comparisons for the metric of per capita publication output.

[1] Accessible at https://databank.worldbank.org/

[2] Accessible at https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

[3] Accessible at https://population.un.org/wpp/

Calculated factors

We calculated a per capita publication output, namely, the number of publications divided by the 2023 World Bank estimate of population (reported in units of thousands under the label "WoS per 1000" in Table 1).

The 4 comparison countries for each factor were later considered in aggregate on a longitudinal basis (see Figure 1); e.g., the Economic comparison group combined the raw publication output of the 4 countries, and did not include Libya. This was chosen over finding a most-similar country because this study does not purport to be an in-depth analysis between two countries; nor did we select synthetic controls because we are not quantifying the "expected" research output of Libya to tease out the effect of a specific event; rather, our RQ1 goal seeks to make simple heuristic comparisons.

WoS Variables

For longitudinal data, we used the "Publication Year" publication counts directly from WoS's online interface for each country or group of countries.

For funding sources, we did an initial assessment of WoS's "Funding Agencies" and found it to be incorrect in spot-checked cases. Thus, we instead examined the Acknowledgement sections of all papers and searched for mentions of "funding"/"funded" or "financial support" or a grant/project number of some kind. Those with such mentions were manually checked for correctness, and then considered to constitute funded projects. We manually coded the resulting papers as having funding sources that were international (INTL), domestic (DOM), or both.

For the affiliations of research teams, we parsed the "Address" field. The majority of the WoS records included unambiguous lists of author-affiliation pairings, even when there were multiple affiliations; the affiliations listed in this field included a mention of their respective country. We normalized each individual author as either having international, domestic, or dual (both international and domestic) affiliation, and counted how many authors of each category were authors on the paper.

Results

WoS publication output comparisons

In Table 1, we compare Libya to other countries along a few heuristic axes: Regional, Economic, Development, and Population. We primarily consider how much research is being produced per capita (or more precisely, per 1,000 population, in the "WoS per 1000" metric).

Among countries with similar population, Libya has a respectable 1.07 publications per 1000 capita – if we exclude the outlier of an economically and developmentally advanced Hong Kong. However, when compared with other North African countries, Libya's publications per capita is second lowest in the region, only ahead of the less-populated Mauritania. Libya is also at the lower end of countries that have similar GDP and countries that have similar HDI scores. Though untested, it appears that the factor of population may be more indicative of potential research output than the other factors.

In the process of selecting comparisons, Turkmenistan arose as a country that is similar in terms of economy, development, and population. Libyan researchers are producing more than 20 times as many articles as Turkmen researchers, per capita. Thus, when considering the full intersectional profile of Libya, research output seems apropos to the context. However, when Libya's research output is compared to similar countries according to a single factor, its relative research performance is poor, corroborating our hypothesis for RQ1.

Table 1. Libyan scientific output from 2004-2024, compared to similar countries. Subtables consider the closest countries in Region (North Africa), GDP (Gross Domestic Poduct; World Bank 2023), HDI (Human Development Index; UNDP 2022), and Population (UN DESA 2023). "WoS" is the number of Web of Science articles produced in the time period listing the country in its "Address" field, while "WoS per 1000" is that value divided by the population in thousands. "h-index" (SCIMAGO 2024) approximates research impact.

	Region	GDP	HDI	Population	WoS	WoS per 1000
Regional comparison						
Mauritania	*North Africa*	$ 10,453	0.540	5,022,000	958	0.19
Morocco	*North Africa*	$ 141,109	0.698	37,713,000	88,276	2.34
Algeria	*North Africa*	$ 239,899	0.745	46,164,000	94,292	2.04
Tunisia	*North Africa*	$ 48,530	0.732	12,200,000	108,976	8.93
Libya	***North Africa***	**$ 50,492**	**0.746**	**7,306,000**	**7,821**	**1.07**
Egypt	*North Africa*	$ 395,926	0.728	114,536,000	333,232	2.91
Economic comparison						
Turkmenistan	Central Asia	*$ 59,877*	0.744	7,364,000	373	0.05
Jordan	Middle East	*$ 50,814*	0.736	11,439,000	74,116	6.48
Libya	**North Africa**	***$ 50,492***	**0.746**	**7,306,000**	**7,821**	**1.07**
Uganda	East Africa	*$ 49,273*	0.550	48,657,000	28,005	0.58
Tunisia	North Africa	*$ 48,530*	0.732	12,200,000	108,976	8.93
Development comparison						
Brazil	South	$	*0.760*	211,141,000	1,170,519	5.54
Colombia	South	$ 363,540	*0.758*	52,321,000	162,943	3.11
Libya	**North Africa**	**$ 50,492**	***0.746***	**7,306,000**	**7,821**	**1.07**
Algeria	North Africa	$ 239,899	*0.745*	46,164,000	94,292	2.04
Population comparison						
Hong Kong	East Asia	$ 382,055	0.956	*7,443,000*	400,430	53.80
Turkmenistan	Central Asia	$ 59,877	0.744	*7,364,000*	373	0.05
Libya	**North Africa**	**$ 50,492**	**0.746**	***7,306,000***	**7,821**	**1.07**
Kyrgyzstan	Central Asia	$ 13,988	0.701	*7,074,000*	4,456	0.63
Paraguay	South	$ 42,956	0.731	*6,844,000*	7,019	1.03

Longitudinal growth in WoS publication output

In Figure 1, we plot the percent growth in number of yearly WoS publications since 2004. Whereas the WoS as a whole showed a global trend of increasing research output (153% increase from 2004 to 2023, the last complete year in our study), Libya showed a much more marked increase (817%) during the same period. Libya's overall publication growth far outstrips that of population-matched countries (285%

growth), HDI-matched countries (366% growth), GDP-matched countries (652% growth).

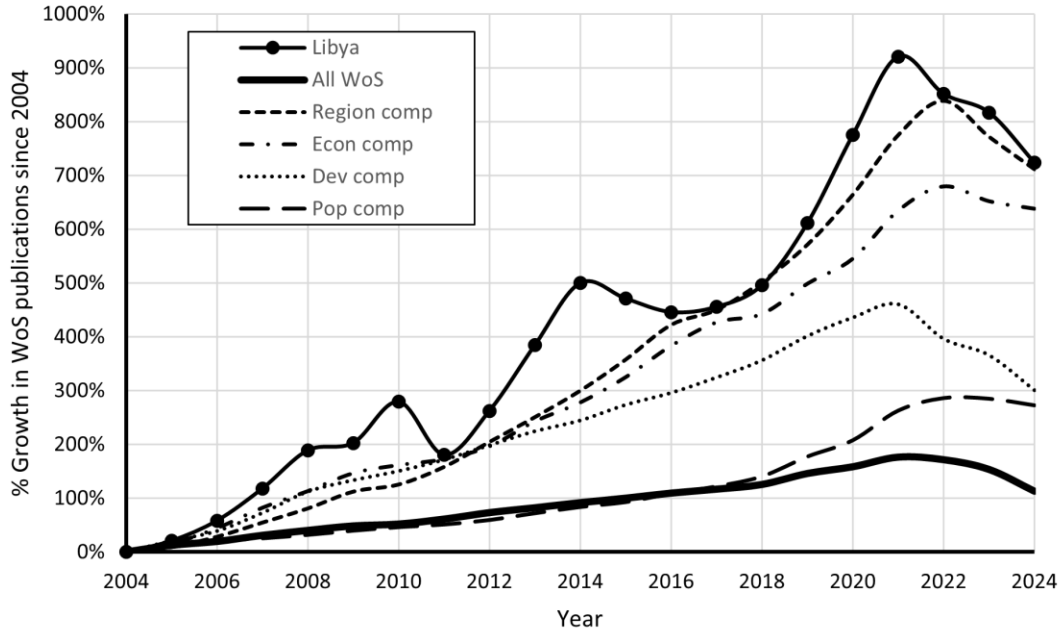


Figure 1. Libya's percent growth in publications over time, compared to regional, economic, development, and population comparison groups. Percentages are calculated relative to research output in 2004. Some salient events over this timeline include a Libyan revolution in 2011, an ongoing civil war since 2014, armed conflict in 2019, and COVID-19 in 2020.

Libya's trend is similar to that of the North African region (772% growth), This growth was achieved despite large dips in the rate of publication growth, concurrent with the February 2011 revolution and the beginning of the 2014-2015 civil war (see these years in Figure 1). Interestingly, the rest of North Africa, which also experienced the Arab Spring in 2011, did not demonstrate as severe a drop in publication growth during that event. Of course, the Libyan civil war is localized to Libya and its lasting effects were not visible elsewhere in the region or the world.

This contradicts our hypothesis for RQ2 that Libyan research is relatively unstable. While a negative effect was visible during periods of national turmoil, the increase of publications has continued at pace with the North Africa region, or better.

Table 2. Funding sources acknowledged in Libyan Web of Science publications from 2004-2024 (WoS Publ).

	WoS Publ	%
Unfunded	5,683	73%
Funded	2,138	27%
International	1,768	23%
Joint	185	2%
Domestic	185	2%
Total (Libya)	7,821	100%

Funding sources for Libyan WoS publications

Focusing on the acknowledged funding sources in the 7,821 Libyan WoS publications, Table 2 shows that 73% of publications were unfunded. Of the publications with funding, 83% received their funding exclusively from outside of Libya. Libyan funding was only acknowledged in 370 publications (4% of the total). Considering funding sources over time, Figure 2 displays the raw count of publications that acknowledge domestic vs. international vs. joint domestic-and-international funding sources. Interestingly, the majority of the growth in WoS publications has occurred in unfunded work. If we considered only unfunded work (white area in Figure 2), there would still be a 546% increase in research output from 2004 to 2023.

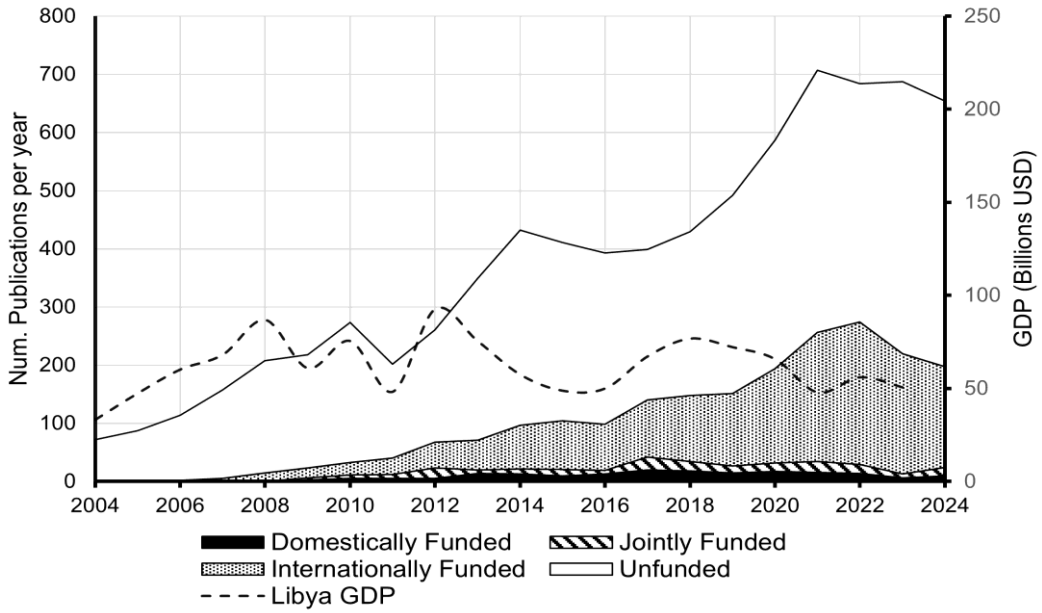


Figure 2. Number of Libyan WoS publications acknowledging domestic vs. international funding sources, yearly, since 2004. Funding sources are stacked (e.g., of 707 publications in 2021, 451 were unfunded, 222 internationally funded, 16 jointly funded, and 18 domestically funded), and the GDP for the same time period is overlaid for comparison.

Over the 20-year period, there is consistent growth in the international funding, with an additional bump in 2021 and 2022, concurrent with global trends of COVID-19 funding. However, there was little growth in the domestic funding, and even in joint funding between international and domestic sources.

It is also evident that unfunded publications bear the brunt of the effect of national-scale events such as the 2011 revolution and the 2014-2015 civil war – financially supported articles show little, if any, effect of those tumultuous events. We have overlaid GDP information over the figure, which fluctuated during the study period and during those events. This general economic indicator fluctuates widely and seems to have had no direct impact on the research output of Libya, as it does not

Table 3. Composition of authorship teams in Libyan WoS publications from 2004-2024 (WoS Publ). "Libyan – with Diaspora" indicates the presence of an author with a dual affiliation, one domestic and another international.

	WoS Publ	% of Publ
Libyan - Local only	1,304	16.7%
Small team (< 5)	1,112	14.2%
Large team (≥ 5)	192	2.5%
Libyan - with Diaspora	75	1.0%
Small team (< 5)	65	0.8%
Large team (≥ 5)	10	0.1%
International - Libyan Majority	762	9.7%
Small team (< 5)	403	5.2%
Large team (≥ 5)	359	4.6%
International - Foreign Majority	5,190	66.4%
Small team (< 5)	1,880	24.0%
Large team (≥ 5)	3,310	42.3%
Missing data	490	6.3%
Total (Libya)	7,821	100.0%

display a similar trend of growth (unfunded and internationally funded work) or stability (domestically funded or jointly funded work).

While this data does show that Libya's research sector is under-funded (RQ3a), it also shows that research in Libya continues to grow despite the under-funding.

WoS publication co-authorship team composition

Recognizing that there is international involvement in the Libyan research sector, Table 3 and Figure 3 consider the composition of co-authorship teams for Libyan WoS publications.

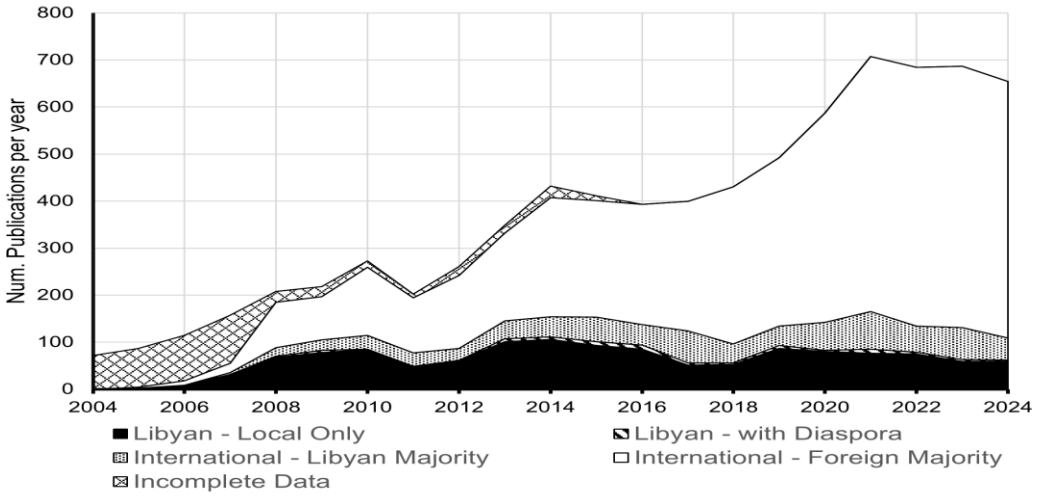


Figure 3. Number of Libyan WoS publications with teams of domestic Libyan authors vs. international authors, yearly, since 2004. Each paper includes at least one Libyan author. "Libyan – with Diaspora" includes Libyans who have a dual affiliation between a Libyan institution and a foreign institution. A division between 50% of the team being Libyan "International – Libyan Majority" or "International – Foreign Majority" indicates that over 50% of the team was either Libyan or non-Libyan. "Incomplete Data" is comprised of 490 studies whose "Address" fields could not be uniquely disambiguated within WoS.

In particular, we classified each author as one who listed a Libyan affiliation (assumed to be Libyans), an international affiliation (assumed to be foreigners), or both (assumed to be Libyan diaspora, namely, those who are working or studying abroad). If a team of co-authors included foreign authors and was more than 50% foreign, we considered the team an "Foreign Majority" team; if fewer than 50% were foreign, we considered it an "Libyan Majority" team. Note that the WoS did not track unambiguous information for this analysis in some of its older articles (6.3%) and these articles were excluded from the classification in this section.

In Table 3 we see that the majority of research from Libya was done in collaboration with foreign-majority authorial teams (66.4%), and oftentimes on large collaborations with over 5 authors (42.3%). However, publications with only Libyan authors in Libya also made contributions (16.7%), but in contrast to the foreign-majority teams, these seemed to focus on smaller teams of fewer than 5 authors (14.2%).

In Figure 3, we can see that publications from Libyan-only teams increased until about 2010 but has not grown much since then. Instead, it was predominantly international partnerships – whether Libyan Majority or Foreign Majority – that accounted for the large rise in publications over the last 20 years. This means that other factors were less explanatory for growth, for example, a 2012 initiative by Libyan authorities to require that scholarship recipients in mobility programs should list a dual Libyan affiliation alongside their foreign affiliation. The corresponding "Libyan – with Diaspora" is poorly represented and it is hard to see a large increase in WoS publication output.

The effect of political turmoil in the 2011 revolution and 2014-2015 civil war was more pronounced for Local-only Libyan teams, though it is also present for other types of teams.

Discussion

Rapid growth in Libyan research productivity, amidst political and societal disruption

The main result of this study is that there was rapid growth in Libyan scientific productivity over the last 20 years, a trajectory which has been heretofore undocumented. The strong growth rate in the North Africa region as a whole was previously only documented for the early part of our study period with publication records up until 2012 or 2013 (Landini et al., 2015; Medina, 2015), and a 7-fold increase in the wider region of the African continent was reported between 2004 and 2019 (Ali & Elbadawy, 2021).

Unfortunately, many studies on bibliometric trends in the region often excluded Libya from analyses due to its low research output (Aggarwal et al., 2020; Ali & Elbadawy, 2021) or failed to select it even among the set of North African countries (Landini et al., 2015; Medina, 2015; Radwan, 2018), or focused on a particular field of study rather than on the productivity of individual countries (Chaabna et al., 2021). However, Siddiqi et al. give a highly relevant and thorough treatment of the Middle East-North Africa (MENA) region according to productivity, indigeneity, and

specialty of the countries' scientific output. Libya and many other MENA countries were shown to increase in global share of publications over the time period of their study; however, its analysis is on older WoS data from 1981-2013 (Siddiqi et al., 2016), which excludes some events of crucial interest for our RQ2, and the issues of RQ3 are also unaddressed for Libya.

The robust growth in research publications in the midst of political turmoil in Libya contradicts our RQ2 hypothesis that Libya's long periods of political instability would correspond with impairment of scientific productivity. While the growth rate decreased slightly during and shortly after events such as the 2011 revolution, the long-term trajectory of growth continued; scientific growth was altogether unaffected by armed East vs. West conflict in 2019. (Note, however, that a global post-COVID decline in 2022-2023 was indeed reflected in the number of Libyan research publications.) This is all the more noteworthy given the expectation that political stability is a precondition for vibrant NISs (Allard et al., 2012; Feng, 1997; Siddiqi et al., 2016), that war and conflict were the greatest challenge facing Libyan universities (UNIGOV, 2016), and the anecdotal evidence from surveys and site visits in the IBTIKAR project considered the instability a barrier (UNIMED, 2024). Relative to unfunded publications, funded publications were less negatively impacted by the political turmoil. Something similar can be said about international-majority teams. We postulate that unfunded, domestic research work depended on societal structures that were affected by armed conflict, whereas funded projects and internationally collaborative research work had a level of invested infrastructure that was less quickly destroyed, and hence less volatile in turmoil.

Libya's trend of productivity growth was observed amidst a fluctuating economy and low reported levels of funding. Though oil and gas output from Libya was unstable through our study period, and the GDP correspondingly, this appears to have no effect on research output. Though UNESCO's Institute for Statistics does not have statistics on Research & Development expenditures for Libya, we surmise that scientific productivity is uncorrelated with GDP because little of the GDP is allocated for research activities.

While there was growth in scientific productivity, the volume of publications from Libya remains on the lower end of regional, economic, and human development comparisons, validating our hypothesis for RQ1. It is most similar in research productivity to countries of similar population (Table 1) and even compares favorably with most of them. Our heuristic selection of factors, and the publication patterns within them, suggest that a low population size may limit the research capabilities of Low-to-Middle Income Countries.

International partnerships and capacity-building funding

One clear result of our work is that most of the growth in Libyan research has involved international-majority authorship teams, a partial answer to RQ3. This effect was previously reported a decade ago as a decrease in "indigeneity" (Landini et al., 2015) of research in Libya and many other parts of the MENA region. Despite some difference in definitions (they only mentioned the address of the contact author, instead of the percent composition of author affiliations), we assert that, indigeneity

continues to decrease in Libya as the overall productivity increases, following the trend observed by Landini et al.

Note however, that this does not necessarily mean that international funding is directly responsible for research work in Libya, since "unfunded" publications showed great increases alongside "internationally funded" publications during the time period of our study. We suggest that the role of international funding to date has not been direct research support that would have been named in WoS papers' Acknowledgments sections, but rather an indirect capacity-building investment in Libya-International research collaborations, and the structures that allow them to occur. Foreign funding of this kind has primarily originated from Europe, including mobility programs, UNIGOV, Libya Restart, and IBTIKAR. International funding has thus made cooperation between Libya and international entities more possible.

International partnerships are also likely to arise out of mobility programs from Libya, as Libyan graduates maintain relationships with the institutions at which they studied. We suspect many not-precise-enough statements mentioned in Acknowledgments sections (e.g., "Embassy of Libya in Malaysia" for "supporting this research") actually had financial support provided by Libyan government-sponsored higher education mobility programs. Although research is part of postgraduate studies for mobility, the achievement in mobility programs tends to be a diploma rather than the research that it took to get that degree. This mindset potentially explains why mobility programs were rarely mentioned in Acknowledgments.

Limitations

Research written in Arabic, particularly in the humanities, often goes unrecognized by global platforms like the WoS due to language barriers, limited access to international publishing, and insufficient institutional support for translation and dissemination. Despite its rich intellectual contributions in fields like literature, history, and philosophy, Arabic research remains underrepresented globally. Bridging this gap requires initiatives such as promoting translations, fostering international collaborations, and creating platforms to highlight Arabic scholarship, ensuring these valuable works gain the recognition they deserve.

Practically speaking, picking periods to calculate percent growth of WoS publications is inherently noisy. Thus, comparisons of percent growth are approximate. The North African region, for example, exhibits a similar growth trend to Libya, and the start and endpoint of the percent growth comparison will dictate whether the country or region exhibits larger growth.

Also, it was inherently difficult to determine the funding status and the funding sources of papers by their Acknowledgments section alone. Noting that the calculated WoS fields were not fully accurate, we attempted our own computationally assisted manual review. However, we still speculate that the actual funding rate is higher and that some systematic biases have prevented more attribution of funding. In particular, funding from the Libyan government through mobility programs was likely underrepresented, given that the Libyan government-

sponsored higher education mobility programs did not obligate the grantees to acknowledge their financial support in Acknowledgments sections.

The Libyan innovation system and Future work

The potential of Libyan innovation is far greater than the current domestic NIS is able to support. The rapid growth in number of publications with international teams demonstrates this potential – it would not be possible if Libyan researchers were entirely lacking the intellectual capital necessary to carry out top-tier research. Thus, in the presence of international team and funding structures, Libyan researchers have been able to sustain rapid growth.

Rather than looking exclusively to the international collaborations and investments that led to the current growth, we may also ask what other types of domestic policies, programs, or other actions can take advantage of the under-utilized research sector in Libya. These initiatives can be informed by further work in innovation studies in Libya. For example, Libya Restart and IBTIKAR projects (UNIMED, 2020, 2024) noted that there are internal struggles with a lack of research administration and funding or projects. Namely, in Libya's NIS, there is no reliable structure for funded research projects. International partnerships provide this type of administrative structure, and it would be instructive to consider what other types of administrative structures would be able to tap into the same research capabilities that the international collaborations are currently tapping into.

Continuing our work here, future studies will need to establish the link between scientific productivity and international capacity building actions. This will enable foreign funders to determine their return on investment, and will also provide a guide for any potential domestic investment in research by the Libyan government. Similarly, Libyan-sponsored mobility programs should be further analyzed to establish how they have impacted scientific productivity. This will enable Libyans to evaluate the benefits of popular programs and compare it with potential domestic investments.

More substantially, although literature on NISs in the region often leaves out Libya (Djeflat, 2004), our results demonstrate that international influence is a key component of the current NIS in Libya. Future work can more precisely identify the players in the trans-national aspects of Libya's innovation system in order to develop policies for encouraging R&I. As new domestic policy actions are taken towards innovation, further studies will need to address underlying internal barriers to having an effective NIS, such as weak interactions between actors (Hamidi & Benabdeljalil, 2013).

Conclusion

A retrospective bibliometric study of Libya's Web of Science publication productivity has shown robust growth over the last 20 years, even through political turmoil and despite lack of funding. International partnerships are noted as a key correlate of this growth, perhaps supported by capacity building projects and mobility programs. While the overall scientific output from Libya is currently low relative to regional, economic, and developmental comparisons, the growth also

suggests existence of substantial intellectual capital that could sustain expansion in research and innovation.

Acknowledgments

The authors would like to thank Dr. Abdurrauf Gusbi and Mohammed Gusbi for supporting scholarly activities within SHU CRI, and Abdulmalek Baitulmal for discussions on innovation in Libya. Illustratively, no funding was received for conducting this study.

References

Aggarwal, A., Patel, P., Lewison, G., Ekzayez, A., Coutts, A., Fouad, F. M., Shamieh, O., Giacaman, R., Kutluk, T., Khalek, R. A., Lawler, M., Boyle, P., Sarfati, D., & Sullivan, R. (2020). The Profile of Non-Communicable Disease (NCD) research in the Middle East and North Africa (MENA) region: Analyzing the NCD burden, research outputs and international research collaboration. *PLoS ONE*, *15*(4), 1–19. https://doi.org/10.1371/journal.pone.0232077

Ali, W., & Elbadawy, A. (2021). Research output of the top 10 African countries : An analytical study. *COLLNET Journal of Scientometrics and Information Management*, *15*(1), 9–25. https://doi.org/10.1080/09737766.2021.1934181

Allard, G., Martinez, C. A., & Williams, C. (2012). Political instability, pro-business market reforms and their impacts on national innovation systems. *Research Policy*, *41*(3), 638–651. https://doi.org/10.1016/j.respol.2011.12.005

Chaabna, K., Cheema, S., Abraham, A., Maisonneuve, P., Lowenfels, A. B., & Mamtani, R. (2021). The state of population health research performance in the Middle East and North Africa: a meta-research study. *Systematic Reviews*, *10*(1), 1–12. https://doi.org/10.1186/s13643-020-01552-x

Djeflat, A. (2004). Knowledge economy for the MENA region. In *World Bank Institute Report*. https://documents1.worldbank.org/curated/en/615321468051039920/pdf/502690WP0Innovation0Box0342042B01PUBLIC1.pdf

Elgamri, A., Mohammed, Z., El-Rhazi, K., Shahrouri, M., Ahram, M., Al-Abbas, A. M., & Silverman, H. (2024). Challenges facing Arab researchers in conducting and publishing scientific research: a qualitative interview study. *Research Ethics*, *20*(2), 331–362. https://doi.org/10.1177/17470161231214636

Feng, Y. (1997). Democracy, political stability and economic growth. *British Journal of Political Science*, *27*(3), 391–418.

Globerman, S., & Shapiro, D. (2003). Governance infrastructure and US foreign direct investment. *Journal of International Business Studies*, *34*, 19–39.

Hamidi, S., & Benabdeljalil, N. (2013). National Innovation Systems: The Moroccan Case. *Procedia - Social and Behavioral Sciences*, *75*, 119–128. https://doi.org/10.1016/j.sbspro.2013.04.014

Landini, F., Malerba, F., & Mavilia, R. (2015). The structure and dynamics of networks of scientific collaborations in Northern Africa. *Scientometrics*, *105*(3), 1787–1807. https://doi.org/10.1007/s11192-015-1635-1

Leydesdorff, L., & Meyer, M. (2006). Triple Helix indicators of knowledge-based innovation systems. Introduction to the special issue. *Research Policy*, *35*(10), 1441–1449. https://doi.org/10.1016/j.respol.2006.09.016

Lundvall, B. A., Johnson, B., Andersen, E. S., & Dalum, B. (2002). National systems of production, innovation, and competence-building. *Research Policy*, *31*(2), 213–231. https://doi.org/10.1016/S0048-7333(01)00137-8

Medina, F. (2015). The output of researchers in Morocco compared to some North African countries from 1996 to 2012, and its relationship to governmental major decisions on higher education and scientific research. *Scientometrics*, *105*(1), 367–384. https://doi.org/10.1007/s11192-015-1701-8

Radwan, A. (2018). Science and innovation policies in north African countries: Exploring challenges and opportunities. *Entrepreneurship and Sustainability Issues*, *6*(1), 268–282. https://doi.org/10.9770/jesi.2018.6.1(17)

Siddiqi, A., Stoppani, J., Anadon, L. D., & Narayanamurti, V. (2016). Scientific wealth in middle east and North Africa: Productivity, indigeneity, and specialty in 1981-2013. *PLoS ONE*, *11*(11), 1–19. https://doi.org/10.1371/journal.pone.0164500

UNIGOV. (2016). *Exploring the Challenges for Higher Education in Libya*. (Tempus Project Ref. Number 530720). Modernizing University Governance and Management in Libya (UNIGOV). https://tempus-unigov.old.ogpi.ua.es/sites/default/files/Exploring the challenges %28web%29_0.pdf

UNIMED. (2020). *Libya Restart: A Journey Analysis*. Mediterranean Universities Union (UNIMED), Rome. https://www.uni-med.net/wp-content/uploads/2020/01/Libya-Restart-UNIMED.pdf

UNIMED. (2024). *IBTIKAR: Promoting research and innovation environment in the Libyan higher education system*. Mediterranean Universities Union (UNIMED), Rome. https://ibtikarproject.eu/

https://doi.org/10.51408/issi2025_040

Research and Application on Multiple Topic Association Fusion Method Based on Neural Network and Evidence Theory

Shuying Li[1], Xian Zhang[2], Jiahui Li[3], Xin Zhang[4], Haiyun Xu[5]

[1]*lisy@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[2] *zhangx@clas.ac.cn,*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[3]*Lijh@las.mail.ac.cn*
University of Chinese Academy of Sciences (UCAS), Beijing 321000 (China)

[4] *zhangx@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[5]*xuhaiyunnemo@gmail.com*
Shandong University of Technology, Zibo City, 255000 (China)

Abstract

This paper contributes to the field of multivariate theme association analysis by proposing a novel data fusion method for patent text theme analysis. The method leverages multiple theme data association features from patent text mining. The methodology of the study involves the extraction of three thematic correlations: term co-occurrence, citation- term coupling, and patent assignee-term coupling. Corresponding matrixes are then constructed, thereby facilitating the analysis of the data. A neural network and evidence-based fusion method is then developed to generate matrixes that are enhanced with information and integrate multi-source uncertainties. Empirical validation using graphene sensing patents demonstrates the method's effectiveness, revealing complementary thematic features and enhanced information richness in fused matrix. The results reveal significant differences among the three types of thematic correlations, highlighting their complementary nature in revealing thematic features. The fused matrices exhibit enhanced information richness and reduced dispersion, effectively capturing both dominant and rare thematic associations. This study underscores the potential of the proposed method to provide comprehensive and precise tools for thematic analysis in patent texts.

Introduction

In the domain of information fusion and knowledge mining, research on multi-topic correlation fusion methods is of considerable theoretical importance and practical application value. With the rapid development of information technology and the increasing demand for multi-source data integration, the efficient extraction, integration, and analysis of multi-topic correlation information from vast amounts of patent literature have become critical for driving technological innovation, guiding industrial upgrading, and predicting technological trends. In light of the rapid advancements in big data and artificial intelligence technologies, patent literature

analysis is undergoing a transformation toward greater complexity and systematization, signifying a significant shift in the research paradigm.

The core of multi-topic correlation fusion identification methods lies in the integration of information from different topic correlations and multiple objectives. This method enables the synthesis of evidence from multiple uncertain information sources to construct an information enhancement matrix that is rich in topic correlations. Consequently, it effectively addresses the limitations of single-relationship types and more accurately reflects the similarity between topics. The fundamental principle underlying this approach is the extraction of consistent information from multi-topic correlations, with the objective of addressing uncertainties caused by various factors, including domain-specific topic terms, cross-topic terms, emerging topic terms, and high-frequency topic terms. Evidence theory, as an efficient information fusion technology, has the ability to clearly distinguish between unknown and uncertain information, and is able to realize the deep integration of information in multiple dimensions(Xiao, 2023). Its powerful fault-tolerant mechanism provides a solid theoretical foundation for the information fusion process(Pan et al., 2021). However, when faced with conflicting evidence, the theory of evidence may encounter limitations, which in turn affects the accuracy of the final judgment(Hamda et al., 2023). In contrast, neural network algorithms demonstrate considerable advantages in the domain of multi-topic relevance fusion recognition due to their superior fault-tolerance performance, efficient hierarchical processing capability, powerful self-learning ability, flexible adaptability, and efficient parallel processing capability. These properties position neural network algorithms as a potent instrument for addressing complex information fusion problems.

In light of the aforementioned analysis, this study proposes a novel integration of evidence theory and neural network algorithm, with the objective of enhancing the processing of relational data in patent documents. The proposed framework encompasses a multifaceted approach, addressing critical aspects such as data source processing, feature-level fusion, decision-level fusion, and data-level fusion. A comprehensive evaluation of conflicting evidence information is achieved through the framework's application, resulting in several notable optimizations. Primarily, the framework enhances the weight allocation for correctly identified evidence. Secondly, it effectively reduces the impact of ambiguous evidence and outlier evidence deviating from the overall level. Finally, it significantly improves the precision and reliability of the system. This innovative approach provides a new technical pathway for patent analysis and offers valuable insights for research in the field of multiple topic association fusion.

Literature Review

In patent analysis, the multi-topic correlation fusion method has shown a wide range of application potential. For example, in technological innovation assessment, this method can accurately identify key technologies and core patents (Huailan Liu et al., 2022); in industrial competition analysis, it helps to reveal competitors' technology layout and market strategies (Song et al., 2023); in policy making, it provides the government with scientific technology trend forecasts and industrial development

suggestions (Yan et al., 2024). In the future, with the continuous development of technologies such as big data and artificial intelligence, the application of multi-topic correlation fusion methods in patent document analysis will be more in-depth and extensive.

Multi-Source Information Fusion

Multi-Source Information Fusion (MSIF) is a comprehensive interdisciplinary field involving multiple disciplines and technologies. In recent years, MSIF has made significant progress in theory and application, but it still faces some key issues and challenges, such as information processing and fusion system design, fusion model and method classification, etc. The application fields of MSIF include but are not limited to military, meteorology, medical, transportation, etc. Tan et al.(2022) designed a multi-source fusion positioning and navigation algorithm based on adaptive filters to integrate the advantages of multiple sensors and provide high-precision and high-reliability positioning and navigation services. Zhu et al.(2024) fused the rolling multi-source heterogeneous information of wind turbines and combined it with the improved PCR6 method to enhance the recognition performance of Rolling Bearing Fault Diagnosis. Li et al.(2024) introduced a multi-source object association method in the radar camera fusion scheme, which significantly improved the vehicle detection accuracy under various adverse conditions and achieved accurate traffic parameter estimation.

In the field of information science, multi-source information fusion can make full use of different information features and internal relationships to achieve information dimension reduction, information integration, information unification, and reduce information uncertainty. Zhang and Lin (2025) proposed a data fusion hierarchical framework that adapts to multi-source and multi-scale schemes, using information gain to aggregate heterogeneous data sources and refine data sets, improving the robustness and effectiveness in processing complex multi-source and multi-scale data environments. Qian et al.(2023) used a variety of generalized multi-granularity rough set models to fuse and utilize multi-source information from multiple perspectives, and adaptively obtained the threshold pairs corresponding to the knowledge granularity through a parameter compensation coefficient, making the model more flexible in practical applications and making decisions more reasonable. Lin et al.(2025)used the information fusion enhanced domain adaptive attention network (IF-EDAAN) to reduce potential feature conflicts, and achieved effective extraction and alignment of temporal and spatial features without domain invariance to improve the efficiency of metastasis diagnosis.

Multivariate Relationship Fusion

In patent analysis, multi-source information may come from different information subjects, and the various multi-relationships between different subjects complement a single relationship. Therefore, patent multi-source information fusion focuses on multi-relationship fusion. The multi-relationship fusion method extracts and integrates multiple related relationships in patent documents, such as subject co-occurrence relationships, citation relationships, patent owner cooperation

relationships and other multi-relationship information, to achieve in-depth identification and analysis of patent themes. This method not only helps to reveal the inherent structure and development context of the technology field but also predicts the emergence of technology fusion and emerging fields. For example, Zhang X. et al.(2024) integrated citation connection relationships, subject association relationships and citation motivation relationships, and proposed a main path identification method for multi-relationship fusion, which effectively improved the identification effect of technology evolution paths in the empirical field. Liu et al.(2024) constructed correlation indicators based on the subject citation relationships, subject relationships, content relationships, and cross-relationships between papers, patents and products, thereby identifying the evolution path of quantum communication technology. In addition, the multi-resource integration model based on the theme graph provides a new perspective and method for the visualization and in-depth mining of patent information. For example, Liu et al.(2022) constructed a hierarchical interactive multi-channel graph neural network based on four relationships: high-order interactions, co-occurrence, hierarchy, and technical knowledge flow to achieve technical knowledge flow prediction. Zhai et al.(2023) constructed a knowledge graph of traditional Chinese medicine based on multi-source heterogeneous data, and used deep learning information, string matching, frequency analysis, association rule Apriori algorithm, etc. to assist researchers in conducting innovative research in the field of traditional Chinese medicine.

Application of D-S Evidence Theory

Commonly used multi-source information fusion methods include classical rough set theory, multi-granularity method, evidence theory and information entropy (Xu et al., 2023). Among them, Dempster-Shafer reasoning (D-S evidence theory) has a strong advantage in processing uncertain information. Zhang et al.(2025) introduced the support matrix based on Dempster-Shafer evidence theory, combined with the hierarchical fusion method, and conducted in-depth research on the information fusion strategy of large-scale multi-source data, and verified that the method is both efficient and effective, and has shown excellent performance in information fusion. Li et al.(2024) realized the information fusion of multi-source incomplete mixed data based on conditional information entropy and DS evidence theory, thereby improving the performance of the classification algorithm. Zhang et al.(2024) proposed a new data enhancement method based on hybrid and Dempster-Shafer reasoning, combined with training deep neural networks to complete recognition or classification tasks, to achieve more effective data enhancement effects and further improve the performance of deep neural networks.

In existing research, the DS method focuses on application scenarios such as intelligent decision-making, while the application of information fusion with patent documents is relatively rare. Patent documents often contain ambiguous, incomplete or contradictory information, and the wide applicability and good robustness of the DS method make it unique in patent document analysis. The D-S evidence theory quantifies this uncertain information through Basic Belief Assignment (BBA) and

uses combination rules to achieve effective fusion of multi-source information. This helps to accurately identify technology trends, evaluate technology maturity, and predict potential technology breakthroughs.

Existing research has primarily applied DS approaches to scenarios such as intelligent decision-making. Nevertheless, the utilization of DS approaches for patent semantic fusion remains comparatively limited. The primary research gaps and shortcomings in this domain can be categorized into three aspects: first, the modeling capability of dynamic topic association is limited. The majority of current methodologies are confined to static topic associations, failing to incorporate effective modeling tools for dynamic topic associations that undergo changes over time (e.g., selection of feature words, adjustment of weights, evolution, etc.). The existing methods demonstrate clear limitations when it comes to capturing and predicting dynamic associations. Second, the extraction of semantic information remains inadequate. Most extant research relies on statistical features or shallow semantic information (e.g., term frequency, co-occurrence relationship, etc.), while the ability to mine deep semantic associations (e.g., semantic similarity of technical concepts, citation text coupling associations, patentee text coupling, etc.) is limited. This may result in the exclusion of significant semantic information during the process of topic association fusion, consequently impacting the precision of the fusion outcomes. In addition, the robustness of evidence conflict processing requires enhancement. Patent text information often contains ambiguous, incomplete, or contradictory content, which further complicates the assessment of evidence quality and the fusion of heterogeneous data. The integration of evidence theory and neural network methodologies has been demonstrated to enhance the balance between the robustness of evidence processing and the accuracy of fusion outcomes. However, these prevailing methodologies continue to fall short in this regard. D-S evidence theory quantifies such uncertain information through Basic Belief Assignment (BBA) and utilizes combinatorial rules to achieve effective fusion of multi-source information. Nevertheless, further enhancement of its robustness remains necessary when addressing conflicts among evidence sources.

In addressing the aforementioned deficiencies, the present study has developed a fusion processing framework for relational data found in patent documents, which integrates neural networks with evidence theory. The framework addresses the limitations of existing methods by enhancing the robustness of evidence conflict processing through the reinforcement of weight assignment to correctly identified evidence. Additionally, the framework improves processing capability for heterogeneous data from multiple sources through multi-level fusion modules. Nonetheless, there is a necessity for further exploration and improvement in dynamic topic modeling, cross-domain adaptation, and deep semantic mining. Future research directions could include the integration of techniques such as graph neural networks and knowledge graphs. This integration has the potential to further enhance the performance and applicability of the multivariate topic association fusion method.

Methodology

Research Framework

This research introduces a novel methodology that integrates neural networks with evidence theory to systematically analyze the intricate relationships among multiple topic associations in patent documents. The proposed approach systematically incorporates three critical data dimensions to holistically capture the intricate relationships among topics. Specifically, the analysis encompasses three aspects: (1) the co-occurrence of subject terms, (2) the coupling relationships between citations and subject terms, and (3) the coupling relationships between patent assignees and subject terms. This multidimensional framework effectively overcomes the inherent limitations of traditional co-occurrence analysis, particularly its susceptibility to loss of information, while simultaneously mitigating the identification inaccuracies that frequently arise from irregular citation patterns in conventional citation-based analysis methods.

The proposed methodology establishes a comprehensive data framework (Fig. 1) through the construction of three distinct types of subject association relationships: subject term co-occurrence, citation-subject term coupling, and patent applicant-subject term coupling, along with their corresponding association matrices. The fusion process employs a weighted allocation strategy that prioritizes evidence information with high reliability while reducing the influence of ambiguous or biased evidence. This approach achieves integration across three levels: feature-level, decision-level, and dataset-level. The result is a robust multivariate topic association fusion model.

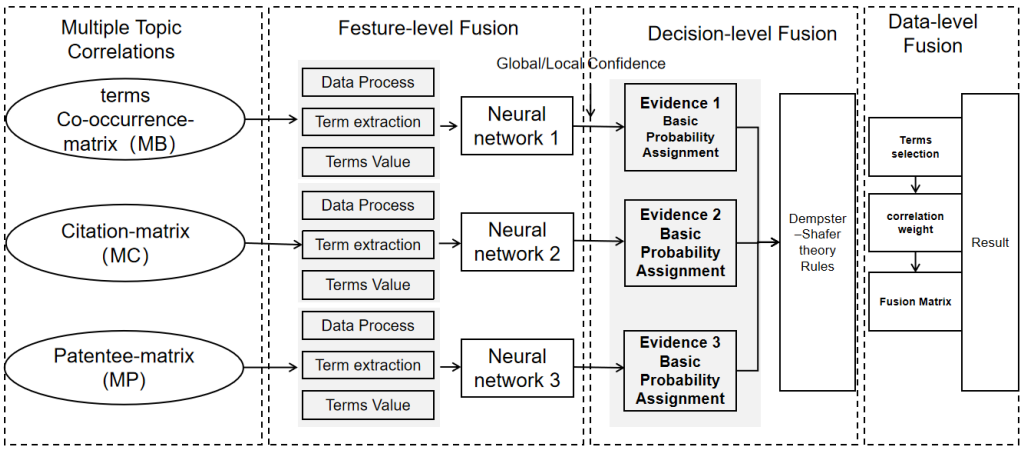


Figure 1. A Neural Network and Evidence Theory-Based Framework for Multi-Topic Association Relationship Integration.

To validate the effectiveness of the proposed method, an empirical study was conducted within the graphene sensing domain as a representative research context. The study successfully demonstrated the integration of the three types of topic association matrices, leveraging the adaptive learning capabilities of neural networks

to refine the fusion process. The incorporation of credibility from diverse evidence sources and the optimization of the fusion mechanism are critical aspects of the proposed method. This framework offers a novel and effective approach to patent text analysis, addressing the limitations of traditional approaches and the complexities of multi-source data integration.

Multiple Topic correlations

Multiple Topic Association Relationships aims to explore in depth how topic terms in patent texts form multiple semantic connections with other measured entities (e.g., patent applicants, citations, and so on). By integrating these different types of associations, we can improve the accuracy and richness of the topic identification process in patent texts. Considering the uniqueness of patent technology innovation activities and patent text characteristics, this study focuses on analyzing diverse subject association relationships in patent texts.

Based on the synergy and inheritance between the subject (e.g., patent applicant), the object (e.g., patent documents), and their characteristics, we have identified three core types of thematic associations by utilizing the information of subject term co-occurrence, citation, and patent cooperation application in patent documents: subject term co-occurrence relationship, citation-subject term coupling relationship, and patent applicant-subject term coupling relationship. The specific definitions of these three relationships are as follows:

(1) Basic association: Terms Co-occurrence Matrix (MB)

This refers to the relationship in which the subject term Ti and the subject term Tj directly co-occur in the same patent document Pm. It reflects the most direct semantic proximity between the subject terms and thus forms the basis for the fusion of multiple relationships in this study.

(2) Extended association: Citation-terms Coupling Matrix (MC)

This relationship indicates that the subject term Ti and the subject term Tj, although distributed in different patent documents, have formed an enhanced association because they are jointly cited by a cited document Ci. The citation of Ci strengthens the association between subject matter Ti and Tj.

(3) Additional Association: Patent Assignee-terms Coupling Matrix (MP)

This refers to the fact that although subject matter Ti and subject matter Tj do not appear in the same patent document or do not have a common patent applicant, a new association path has been formed due to the existence of a cooperative application for patent Pm by their respective corresponding patent applicants Ai and Aj. This relationship connects the originally independent subject terms Ti and Tj through patent applicant Ai, patent document Pm, and patent applicant Aj.

Feature-level integration of multiple thematic correlations

The utilization of feedback mechanisms inherent in neural network algorithms facilitates the implementation of feature-level fusion through the adjustment of weights. This process entails the consideration of diverse combinations of various features, thereby ensuring the effective integration of multi-topic relational data. The fundamental principle underpinning this process is the continuous adjustment of

weights and thresholds within the network through the backward propagation of errors, a process that continues until the sum of squared errors at the output layer of the network is minimized. In this study, the BP neural network model was adopted to preprocess three types of multi-topic relational data, from which representative feature vectors were extracted as inputs for the BP neural network. Based on previous research, four target classifications were identified, namely, domain-specific terms, technology cross-over terms, technology burst terms, and high-frequency terms. Subsequently, eight measurement indicators were selected as the feature values to be fused for these four target classifications, including High-Frequency (HF) (Qaiser and Ali, 2018; Tseng et al., 2007), Term Frequency-inverse Document Frequency (TFIDF) (Chawla et al., 2023), Comprehensively Measure Feature Selection (CMFS) (Yang et al., 2012), Information Gain (IG) (Yu et al., 2022), Term Interdisciplinary index (TI) (Xu et al., 2016), Shannon-Wiener Index (SWI) (Shannon, 1948), Kleinberg burst (KB) (Kleinberg, 2002), and growth rate (GR) (Feng et al., 2020).

A BP neural network model (Fig.2) is constructed for each topic association relationship. The strong generalization and nonlinear mapping capability of the neural network algorithm enables the association of multiple eigenvalues, thereby facilitating the identification and classification of the target type by each topic-associated relationship. The output of the neural network can be utilized as evidence of the efficient utilization of multiple eigenvector changes.

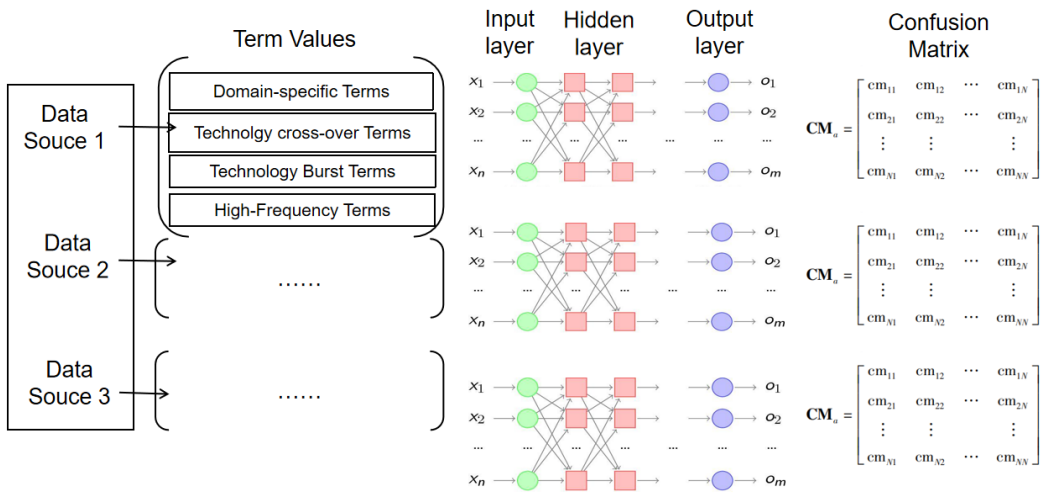


Figure 2. BP neural network model.

The confusion matrix, a statistical tool used for understanding and interpreting data, is a crucial component of this analysis. It delineates the ability of each subject association to recognize the target, and consequently, the global and local credibility of each subject association is calculated. The local credibility is weighted and fused with a posteriori probability output to construct the basic probability distribution function, providing a comprehensive framework for understanding the relationship between credibility and prediction accuracy.

The Confusion Matrix (CM) developed from the BP neural network model classification indicates that the recognition capability of each association relationship between topics varies. These associations include domain feature topic words, technology cross-topic words, technology burst topic words, and high frequency topic words, among others. Theoretically, topic word co-occurrence, serving as the base relationship, exhibits a potential enhancement in recognition performance for both domain feature topics and high frequency feature topics. Similarly, citation-topic word coupling, operating as the reinforcement relationship, is predicted to demonstrate an improved recognition capability for technology burst features. Furthermore, patentee-topic word coupling, functioning as the additional relationship, is expected to show enhanced performance in recognizing technology burst features. It is conceivable that the citation-topic word coupling as a reinforcement relationship would be more efficacious for recognizing technical cross features; alternatively, the patentee-topic word coupling as an additional relationship may be more effective for recognizing technical emergent features. The confusion matrix includes T samples, each containing N distinct target types, with a sample count of Ti (i=1,2,…,N) for each target. The formula is as follows:

$$CM_a = \begin{Bmatrix} \mathrm{cm}_{11} & \mathrm{cm}_{12} & \mathrm{cm}_{1N} \\ \vdots & \vdots & \vdots \\ \mathrm{cm}_{N1} & \mathrm{cm}_{N2} & \mathrm{cm}_{NN} \end{Bmatrix} \quad (1)$$

Where *a* is the number of neural networks; the row subscripts of *cm* in the set of confusion matrices are the true target types; the column subscripts are the target types recognized by the neural network, representing the proportion of the number of samples with target type *i* recognized by the neural network as type *j* to the proportion of samples of type *i*; and the diagonal elements are the percentage of elements of each target type that can be correctly recognized by the neural network.

The BP neural network is employed to evaluate the target recognition classification ability of multivariate topic association by constructing a realistic basic probability assignment function. To this end, it is imperative to calculate the probability that the test sample of class i is classified to class j by the neural network based on the confusion matrix as expressed in equation (2):

$$W_{ij} = cm_{ij} \quad (2)$$

Second, the local credibility of the j^{th} target in the a^{th} neural network is calculated using the following equation（3）：

$$W_{\mathrm{Local}_{\mathrm{aj}}} = cm_{jj} \Big/ \sum_{i=1}^{N} cm_{ij} \quad (3)$$

The global credibility of the neural network is ultimately determined by the following equation(4)：

$$W_{\text{Global}_{aj}} = \sum_{i=1}^{4} Wii \Big/ N \quad (4)$$

Decision-level integration of multiple thematic correlations

Evidence theory, also known as Dempster-Shafer (DS) theory, is a theoretical framework that has found wide application in the fields of multi-source information fusion and decision analysis. The core advantage of this theory lies in its ability to effectively integrate multiple pieces of uncertain information evidence. Through synthesis or reasoning processes, it standardizes and combines information from multi-source data, thereby enhancing the reliability and accuracy of fusion recognition. In this study, "evidence" is defined as uncertain information data involved in target recognition, including domain-specific topic terms, technology crossover topic terms, technology burst topic terms, and high-frequency topic terms. Meanwhile, "DS combination" refers to the process of synthesizing information represented by multi-source data through combination rules. By fusing and reasoning data sources under different topic relationships, it ultimately outputs decision inputs or decision results. The employment of DS evidence theory for decision-level fusion facilitates comprehensive observation of local feature values provided by multi-entity relationships, thereby enhancing the accuracy and reliability of decision-making.

Specifically, the basic probability assignment at the decision level can be achieved based on the global credibility and local credibility of the a^{th} BP neural network, combined with the posterior probability estimates provided by the algorithm classification results. The calculation of the basic probability assignment involves weighting and fusing the local credibility output by the BP neural network with the posterior probability, followed by normalization. $P_{aj'}$ is the output of neural network test samples, and its calculation (5) is as follows:

$$P_{aj}{}' = W_{aj}P_{aj} \Big/ \sum_{i=1}^{4} W_{aj}P_{aj} \quad (5)$$

The basic probability assignment is defined as the sum of all subset likelihood calculations for that S hypothesis to be true, expressing the level of confidence in the event that S is hypothesized to be true, with the formula as in (6):

$$m_a(S_1, S_2, \cdots S_j, \Theta,\ \emptyset) = (W_{Global_{aj}} * P_{a1}',\ W_{Global_{aj}} * P_{a2}',\ \cdots,\ 1 - W_{Global_{aj}},\ 0) \quad (6)$$

Where $\Theta = (S_1,\ S_2,\ \cdots,\ S_j,\ \Phi)$ and S is each hypothesis in the recognition framework for the jth target type, i.e., as in Equation (7):

$$\mathrm{m}\ (\emptyset)\ = 0 \quad (7)$$

,

$$m_a(S_1) + m_a(S_2) + \cdots + m_a(S_j) + m_a(\Theta) = 1$$

The subsequent step involves the utilization of the DS combination rule to adjust the basic probability assignments, which are defined as *n* mutually independent variables. The underlying assumption of this study is that the three relations *MB*, *MC*, and *MP* possess three trust functions within the same identification framework. It is further postulated that *m1*, *m2*, *m3*, and *m4* represent the fundamental probability assignments of their respective features. The combination rule, as in equation 8, is then employed to derive the subsequent results.

$$\mathrm{m(a)} = \frac{1}{\mathrm{K}} \sum_{\mathrm{a1} \cap \mathrm{a2} \cap \mathrm{a3} = \mathrm{a}} m_1(\mathrm{a1}) m_2(\mathrm{a2}) m_3(\mathrm{a3}) m_4(\mathrm{a4}) \qquad (8)$$

where K is the normalization factor:

$$\mathrm{K} = \sum_{\mathrm{A1} \cap \mathrm{A2} \cap \mathrm{A3} = \emptyset} m_1(\mathrm{a1}) m_2(\mathrm{a2}) m_3(\mathrm{a3}) m_4(\mathrm{a4}) \qquad (9)$$

Therefore, based on the fused trust decision to derive which category of features the evidence data (subject terms in this study) belongs to, the feature with the greatest trust is selected as the verdict. If $\mathrm{m}(a_1) = max\{m(a_i),\ a_i \in \Theta\}$, then *a1* is the categorization result.

Data-level integration of multiple thematic correlations

The integration of the identification framework, basic probability assignment, and combination rules of Dempster-Shafer (DS) evidence theory results in the calculation of data fusion results. Subsequently, decision criteria are employed to identify different target objects, thereby achieving data fusion and target recognition for multi-topic relationships. Specifically, the first step involves the utilization of a BP neural network to achieve feature-level fusion, which is a process of feature input and decision output. DS evidence theory is implemented to attain decision-level fusion of multi-topic relationships, a process of combining decisions as input and producing decisions as output.

As illustrated in Figure 3, the data-level fusion comprises two stages: In the initial stage of fusion, DS evidence theory is employed to optimize decision-making regarding identification results, screen topic words, and select the feature word list with the highest degree of belief, thereby forming a comprehensive word list post-fusion. In the second stage, the basic association relationship matrix MB, the enhanced association relationship matrix MC, and the newly added association relationship matrix MP are reconstructed using the method described in Section 3.1.3. The BP neural network is then utilized to calculate the global confidence level of the comprehensive word list, which serves as the fusion association weight and becomes the foundation for calculating and realizing the fusion of multi-topic relationships.

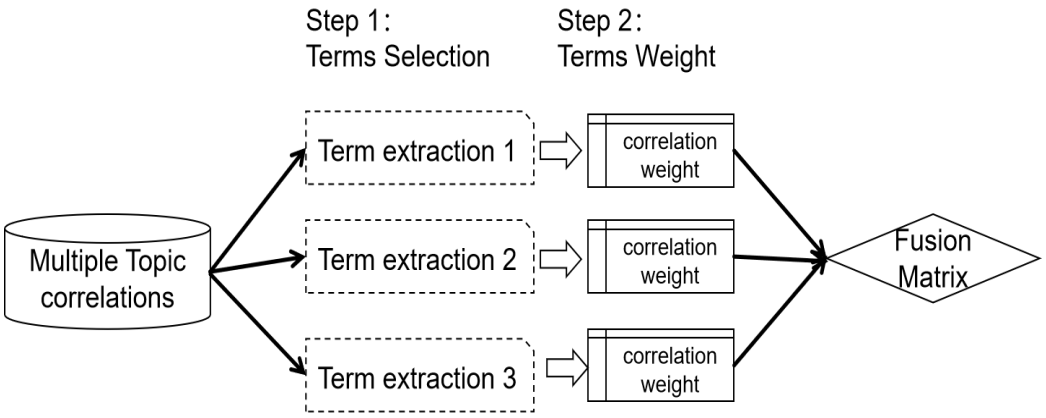


Figure 3. Multi-Relationship Matrix Fusion.

The fusion computation of the multi-relation matrix is intended to amplify the weight of evidence information that is conducive to correct identification, while minimizing the impact of ambiguous evidence information and evidence information that deviates significantly from the overall level. Therefore, this study sets the weights of three neural network relationships based on the global confidence level of the BP neural network, as shown in formulas (10), (11), and (12), respectively. This approach is predicated on the multi-objective characteristic of the feature-level fusion of the three types of matrices.

$$M_{MB} = \frac{W_{\text{Global}_{MB}}}{(W_{\text{Global}_{MB}} + W_{\text{Global}_{MC}} + W_{\text{Global}_{MP}})} \quad (10)$$

$$M_{MC} = \frac{W_{\text{Global}_{MC}}}{(W_{\text{Global}_{MB}} + W_{\text{Global}_{MC}} + W_{\text{Global}_{MP}})} \quad (11)$$

$$M_{MP} = \frac{W_{\text{Global}_{MP}}}{(W_{\text{Global}_{MB}} + W_{\text{Global}_{MC}} + W_{\text{Global}_{MP}})} \quad (12)$$

The weighted weights of the base relationship matrix MB, the enhanced relationship matrix MC, and the added relationship matrix MP are obtained according to the aforementioned method. The relationship fusion is then performed, and the fusion matrix is calculated to obtain the fusion matrix.

Empirical Study

This study selected the field of graphene sensing technology as the empirical research. A comprehensive patent analysis was conducted, yielding a set of subject feature words, including domain-specific terms, technology cross-over terms, technology burst terms, and high-frequency terms. These words were then used to construct three types of multiple topic correlations: terms co-occurrence, citation-

term coupling, and patentee-term coupling. To verify the efficacy of the study's construction method, fusion calculations were performed.

In this study, graphene sensing technology is selected as the empirical technology field of the method due to its highly interdisciplinary nature, encompassing multiple technological domains such as materials, information, and biological sciences. This technology field is also characterized by its dynamism, marked by ongoing research activities, significant innovation, and rapid technological convergence. These characteristics are particularly relevant for the practical evaluation of the method outlined in this paper.

Data preprocessing

This study employs Derwent Innovation as its data source and devises a search strategy by combining the relevant concepts of graphene and sensor technology through Boolean logic. Following the filtration of search results, a collection of 974 patent families' documents was obtained, and the search for data continued up to September 22, 2022.

The feature items that are subsequently extracted from these patent data include subject terms, patent owner, and cited documents.

Of these, the patent owners and cited patent documents are directly extracted from the original data of the search results. The subject terms are extracted from the title and abstract text using text mining methods. The specific process is outlined as follows:

Step 1: The title and abstract text fields of the patent are segmented using the natural language processing (NLP) function of the Derwent Data Analyzer (DDA), resulting in a collection of 20,036 original subject terms (groups).

Step 2: The initial term sets must be rectified. The first step in the process is to convert all texts to lowercase in order to prevent errors. The second step involves the use of a built-in stop word list, thesaurus, etc., in order to remove general stop words, as well as format and grammatical terms in patent documents, DWPI catalog format abbreviations, compound name specifications, British and American spelling specifications, etc. The third step involves the use of Python's NLTK package stop word list to remove meaningless stop words and numbers, merge similar word forms, etc. Experts then perform manual cleaning, merge synonyms, and eliminate general subject words that are not closely related to substantive research, as well as conventional experimental tool names, material names, etc. After the above preprocessing operations, the pre-selected subject word set to be measured is obtained, containing 7873 words.

Step 3: It was implemented to extract feature parameters and feature vector sets. For 7873 pre-selected subject words, four types of features were obtained based on three different network of basic association MB, extended association MC and additional Association MP: namely, domain-specific terms, technology cross-over terms, technology burst terms, and high-frequency terms. Eight measurement indicators were calculated and collected respectively, which were HF, TFIDF, CMFS, IG, TI, SWI, KB and GR. The 8 parameters after standardization were used as indicators to be fused. The data examples are shown in Table 1, Table 2 and Table 3. The 7873

keywords in each type of network are divided into 5511 training samples and 2362 test samples in a ratio of 7:3.

Table 1. Top 20 Terms Eigenvalues in basic relation MB network.

Terms	*Frequency Features*		*Domain Feature*		*Interdisciplinary Feature*		*Breakthrough Feature*		*Category*
	HF	*TF-IDF*	*CMFS*	*IG*	*TI*	*SWI*	*KB*	*Max (GR)*	
analyte	0.430	0.040	0.680	0.299	0.340	0.411	0.000	0.015	1
nanopore	0.118	0.125	0.700	0.440	0.072	0.431	0.013	0.029	1
binding member	0.067	0.929	0.767	0.667	0.007	0.154	0.000	0.000	1
response data	0.040	0.550	0.692	0.557	0.019	0.433	0.000	0.000	1
gip agonist peptide	0.033	1.000	0.671	0.555	0.006	0.154	0.000	0.000	1
sensing layer	0.020	0.054	0.684	0.486	0.016	0.262	0.000	0.004	2
photoluminescent nanostructure	0.020	0.092	0.716	0.518	0.015	0.212	0.000	0.008	2
ionic liquid	0.019	0.050	0.670	0.371	0.010	0.236	0.000	0.000	2
graphene channel	0.018	0.068	0.658	0.455	0.015	0.253	0.000	0.004	2
balloon	0.017	0.052	0.702	0.470	0.010	0.178	0.000	0.003	2
detecting	0.541	0.010	0.635	0.198	0.766	0.462	0.000	0.041	3
sample	0.385	0.022	0.655	0.232	0.353	0.436	0.000	0.016	3
solution	0.320	0.018	0.681	0.300	0.486	0.413	0.000	0.013	3
material	0.315	0.014	0.654	0.236	0.665	0.598	0.000	0.078	3
antibody	0.299	0.026	0.666	0.200	0.330	0.533	0.000	0.065	3
method	1.000	0.011	0.670	0.182	0.893	0.531	1.000	0.075	4
layer	0.662	0.021	0.701	0.266	0.565	0.495	0.479	0.044	4
sensor	0.651	0.013	0.669	0.249	1.000	0.514	0.687	0.013	4
surface	0.583	0.013	0.658	0.175	0.691	0.533	0.697	0.092	4
patient	0.581	0.018	0.694	0.306	0.743	0.554	0.530	0.011	4

Table 2. Top 20 Terms Eigenvalues in extended relation MC network.

Terms	*Frequency Features*		*Domain Feature*		*Interdisciplinary Feature*		*Breakthrough Feature*		*Category*
	HF	*TF-IDF*	*CMFS*	*IG*	*TI*	*SWI*	*KB*	*Max (GR)*	
analyte	0.685	0.063	0.707	0.257	0.359	0.340	0.000	0.135	1
nutritional substance	0.370	0.078	0.809	0.686	0.285	0.556	0.000	0.005	1
binding member	0.198	0.929	0.791	0.672	0.016	0.157	0.000	0.000	1
nanopore	0.181	0.165	0.715	0.451	0.092	0.302	0.000	0.030	1
gip agonist peptide	0.096	1.000	0.692	0.572	0.013	0.157	0.000	0.000	1

composite material	0.038	0.030	0.692	0.466	0.030	0.225	0.000	0.020	2
electrode array	0.036	0.106	0.754	0.628	0.039	0.273	0.000	0.000	2
sensing device	0.032	0.051	0.682	0.416	0.028	0.144	0.000	0.007	2
enzyme-free glucose sensor	0.030	0.062	0.710	0.662	0.009	0.157	0.000	0.002	2
intermediate body	0.030	0.087	0.742	0.752	0.018	0.268	0.000	0.000	2
method	1.000	0.029	0.711	0.178	1.000	0.453	0.000	0.092	3
layer	0.528	0.038	0.694	0.260	0.496	0.335	0.000	0.016	3
device	0.487	0.029	0.686	0.257	0.676	0.560	0.000	0.061	3
detecting	0.438	0.019	0.640	0.187	0.626	0.449	0.000	0.039	3
sample	0.398	0.039	0.660	0.236	0.380	0.432	0.000	0.015	3
sensor	0.679	0.022	0.689	0.246	0.977	0.491	1.000	0.020	4
surface	0.587	0.026	0.674	0.135	0.797	0.563	0.759	0.053	4
substrate	0.534	0.029	0.680	0.219	0.688	0.534	0.539	0.118	4
binding	0.460	0.056	0.688	0.194	0.372	0.443	0.285	0.063	4
amino acid	0.436	0.220	0.928	0.186	0.387	0.278	0.000	1.000	4

Table 3. Top 20 Terms Eigenvalues in additional relation MP network.

Terms	*Frequency Features*		*Domain Feature*		*Interdisciplinary Feature*		*Breakthrough Feature*		*Cate-gory*
	HF	*TF-IDF*	*CMFS*	*IG*	*TI*	*SWI*	*KB*	*Max (GR)*	
nutritional substance	0.815	0.191	0.889	0.595	0.437	0.649	0.000	0.069	1
analyte	0.556	0.200	0.795	0.350	0.217	0.653	0.000	0.000	1
patient	0.407	0.177	0.872	0.476	0.232	0.544	0.000	0.000	1
target analyte	0.370	0.365	0.794	0.583	0.161	0.594	0.000	0.000	1
nanopore	0.284	1.000	0.842	0.722	0.119	0.758	0.000	0.000	1
cnt	0.049	0.174	0.819	0.747	0.039	0.593	0.000	0.000	2
diabetes	0.037	0.059	0.706	0.350	0.010	0.211	0.000	0.000	2
conductive polymer	0.037	0.130	0.746	0.534	0.021	0.409	0.000	0.000	2
sensor chamber	0.037	0.130	0.744	0.527	0.021	0.409	0.000	0.000	2
pressure sensor	0.037	0.130	0.763	0.595	0.010	0.211	0.000	0.000	2
compound	0.506	0.135	0.768	0.206	0.458	0.824	0.000	0.000	3
graphene	0.457	0.095	0.751	0.212	0.308	0.594	0.000	0.130	3
sensor	0.444	0.054	0.753	0.230	0.510	0.558	0.000	0.042	3
substances	0.383	0.080	0.755	0.339	0.326	0.473	0.000	0.000	3
organoleptic	0.358	0.095	0.774	0.392	0.351	0.455	0.000	0.000	3
method	1.000	0.075	0.736	0.145	1.000	0.711	1.000	1.000	4
device	0.580	0.083	0.744	0.199	0.365	0.636	0.478	0.111	4
system	0.506	0.095	0.761	0.335	0.563	0.731	0.351	0.093	4
detecting	0.494	0.060	0.704	0.174	0.490	0.576	0.569	0.333	4
sample	0.481	0.148	0.752	0.334	0.174	0.510	0.199	0.083	4

The Confusion matrix and target identification results

The application of the aforementioned method entails the substitution of training samples into the neural network to facilitate preliminary target recognition and classification. Subsequent to the execution of this operation, the fundamental probability assignment derived from fuzzy processing is modified. Concurrently, the evidence theory space is formed, integrating the previously obtained data. The confusion matrix corresponding to the multi-topic relationships are obtained according to the classification results of the entire set.

MB confusion matrix:

$$\begin{bmatrix} 0.9333 & 0.0018 & 0 & 0 \\ 0.0333 & 0.9866 & 0.0077 & 0.1316 \\ 0.1 & 0.0018 & 0.9957 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

MC confusion matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.9849 & 0.0138 & 0 \\ 0.1429 & 0.0022 & 0.9842 & 0.3333 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

MP confusion matrix:

$$\begin{bmatrix} 0.9767 & 0.0159 & 0 & 0 \\ 0 & 0.8889 & 0.0464 & 0 \\ 0 & 0.0159 & 0.9404 & 0.5333 \\ 0 & 0 & 0.0066 & 0.9333 \end{bmatrix}$$

The global and local credibility of the multi-topic relationship are calculated using the confusion matrix, as illustrated in Table 4. The global credibility indicates that the classification recognition rate of the basic relationship MB is high, while the classification recognition efficiency of the newly added relationship MP is relatively low. It is evident that misjudgments occur in the recognition and classification of the three topic relationships. Specifically, the basic relationship MB demonstrates a high recognition rate for domain features, cross features, and burst features, while the enhanced relationship MC exhibits a high recognition rate for domain features, cross features, and word frequency features. Notably, the newly added relationship MP shows a high recognition rate for word frequency features, domain features, and cross features. Consequently, each multi-topic relationship exhibits distinct recognition accuracy rates for various indicators, thereby effectively addressing the issue of substantial confidence disparities during target recognition under a single topic relationship.

Table 4. The retrieval strategy for Grapheny Sensing Technology.

BP Neural Network	*Global Credibility*	*Local Credibility*			
		Frequency Features	*Domain Feature*	*Interdisciplinary Feature*	*Breakthrough Feature*
MB	0.937	0.875	0.996	0.9923	0.884
MC	0.902	0.875	0.998	0.986	0.75
MP	0.887	1	0.966	0.947	0.636

The Feature Terms Selection

In order to mitigate the discrepancy between the output of the neural network and the actual classification of feature words, the output results of the BP network are normalized. Subsequently, the fusion evaluation is performed using the evidence theory (see Table 5) to mitigate the impact of uncertain factors. The core subject word list is selected and merged based on the output fusion classification results and combined with expert opinions. After deduplication, the comprehensive subject word list used in this experiment is obtained, which contains 887 core terms (groups).

Table 5. The Top terms of Basic probability distribution function assignment and fusion recognition results.

Terms	*Network*	$m_{Frequency}$ *(S1)*	m_{Domain} *(S2)*	$m_{Interdisciplinary}$ *(S3)*	$m_{Breakthrough}$ *(S4)*	$m_a(\Theta)$	*Identification Results*
arterial pressure	MB	0.000	0.909	0.000	0.000	0.091	S2
	MC	0.000	0.854	0.000	0.000	0.146	S2
	MP	0.000	0.854	0.008	0.007	0.128	S2
conductivity	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.001	0.003	0.846	0.001	0.150	S3
hemorrhage	MB	0.000	0.909	0.000	0.000	0.091	S2
	MC	0.000	0.854	0.000	0.000	0.146	S2
	MP	0.878	0.000	0.000	0.000	0.122	S1
expression	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.001	0.837	0.023	0.000	0.139	S2
	MP	0.046	0.305	0.478	0.020	0.150	S3
synthetic compounds	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.010	0.651	0.151	0.037	0.151	S2
metal	MB	0.000	0.000	0.000	0.909	0.091	S4
	MC	0.036	0.000	0.000	0.768	0.195	S4
	MP	0.000	0.032	0.579	0.224	0.164	S3
cholesterol	MB	0.006	0.000	0.008	0.895	0.091	S4
	MC	0.006	0.000	0.963	0.023	0.008	S3
	MP	0.002	0.031	0.903	0.040	0.025	S3
conditioner	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.001	0.007	0.836	0.004	0.152	S3
nerve cell	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3

	MP	0.005	0.004	0.839	0.002	0.150	S3
	MB	0.000	0.000	0.000	0.909	0.091	S4
polymers	MC	0.025	0.000	0.000	0.782	0.193	S4
	MP	0.001	0.004	0.998	0.001	-0.004	S3
	MB	0.000	0.899	0.010	0.000	0.091	S2
logistic transport	MC	0.001	0.850	0.002	0.000	0.147	S2
	MP	0.874	0.001	0.002	0.000	0.122	S1
	MB	0.000	0.000	0.909	0.000	0.091	S3
stem cell	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.005	0.004	0.839	0.002	0.150	S3
	MB	0.000	0.000	0.010	0.899	0.091	S4
glucose concentrati on	MC	0.080	0.000	0.060	0.662	0.198	S4
	MP	0.002	0.003	0.997	0.001	-0.003	S3
	MB	0.000	0.908	0.001	0.000	0.091	S2
calibration temperature sensor	MC	0.000	0.845	0.008	0.000	0.147	S2
	MP	0.007	0.184	0.646	0.008	0.155	S3
	MB	0.000	0.000	0.909	0.000	0.091	S3
amyotrophic lateral sclerosis	MC	0.000	0.003	0.845	0.000	0.152	S3
	MP	0.004	0.077	0.754	0.011	0.154	S3
	MB	0.000	0.000	0.909	0.000	0.091	S3
high reactivity	MC	0.000	0.023	0.825	0.000	0.152	S3
	MP	0.004	0.021	0.817	0.005	0.152	S3
	MB	0.000	0.000	0.909	0.000	0.091	S3
progression	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.716	0.009	0.146	0.005	0.124	S1
	MB	0.000	0.881	0.027	0.000	0.091	S2
beverage consumptions	MC	0.001	0.846	0.007	0.000	0.146	S2
	MP	0.001	0.750	0.098	0.014	0.138	S2
	MB	0.000	0.000	0.909	0.000	0.091	S3
pressure sensor	MC	0.008	0.013	0.808	0.012	0.159	S3
	MP	0.429	0.443	0.012	0.001	0.115	S2
	MB	0.000	0.000	0.909	0.000	0.091	S3
transmembr ane pore	MC	0.003	0.412	0.438	0.001	0.147	S3
	MP	0.000	0.000	0.851	0.000	0.149	S3
	MB	0.205	0.478	0.174	0.052	0.091	S2
substrate platform	MC	0.155	0.612	0.034	0.012	0.188	S2
	MP	0.926	0.000	0.007	0.000	0.066	S1
	MB	0.003	0.000	0.000	0.906	0.091	S4
transistor	MC	0.001	0.000	0.984	0.014	0.002	S3
	MP	1.047	0.000	0.000	0.000	-0.047	S1
	MB	0.000	0.000	0.908	0.001	0.091	S3
high reliability	MC	0.002	0.495	0.358	0.000	0.145	S2
	MP	0.004	0.077	0.754	0.011	0.154	S3
	MB	0.000	0.000	0.909	0.000	0.091	S3
antibody	MC	0.000	0.000	0.843	0.004	0.153	S3
	MP	0.002	0.079	0.405	0.209	0.305	S3

The Matrix extraction and fusion calculation

Based on the 887 terms (groups) in the comprehensive term list, the terms co-occurrence, the citation-terms coupling, and patent assignee-terms coupling were

extracted, and 84,958 groups of multivariate relationships were obtained. Based on these relationships, three types of association matrix were calculated and constructed: MB, MC, MP.

The 887 comprehensive subject terms (groups) were substituted into the neural network for target classification. According to the classification results, the MB_{core}, MC_{core}, and MP_{core} confusion matrix corresponding to the multiple topic associations were obtained, and the global credibility and local confidence were calculated based on the confusion matrix. Therefore, the fusion weights of the multivariate subject association network were judged to be 0.361, 0.333, and 0.305, respectively (Table 6).

MB_{core} confusion matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.9767 & 0 & 0.0556 \\ 0.0833 & 0.0233 & 0.9227 & 0.7222 \\ 0.0833 & 0 & 0 & 0.9444 \end{bmatrix}$$

MC_{core} confusion matrix:

$$\begin{bmatrix} 0.8333 & 0 & 0.0065 & 0 \\ 0.1667 & 0.8519 & 0.013 & 0.0909 \\ 0 & 0.2222 & 0.9351 & 0.3636 \\ 0.1667 & 0 & 0.0065 & 0.8182 \end{bmatrix}$$

MP_{core} confusion matrix:

$$\begin{bmatrix} 0.8333 & 0 & 0.0143 & 0.1111 \\ 0.1667 & 0.625 & 0.0143 & 0 \\ 0.3333 & 0.375 & 0.8429 & 0.4444 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Table 6. BP neural network core keyword target recognition results.

BP Neural Network	*Global Credibility*	*Local Credibility*				*Fusion Weight*
		Frequency Features	*Domain Feature*	*Interdisciplinary Feature*	*Breakthrough Feature*	
MB	0.846	0.857	0.977	1	0.548	0.361
MC	0.781	0.714	0.793	0.973	0.643	0.333
MP	0.715	0.625	0.625	0.967	0.643	0.305

Discussion

Theme correlation matrix fusion effect

According to the ranking of the fusion matrix value results, the top 50 terms(group) associations were selected, and their situations in the fusion matrix and the three types of subject association matrices were analysed. Table 7 shows the 50 groups of subject associations with the highest fusion weights. Among them, the top three groups of subject associations with the highest fusion association weights are

(method, solution), (method, patient) and (method, detecting), and their fusion matrix association weights are 0.361, 0.359 and 0.35 respectively. Among them, the subject word with the strongest co-occurrence association is (method, solution), the citation-subject word coupling association with the strongest association is (semiconductor, multi-walled carbon), and the patent applicant-subject word coupling association with the strongest association is (resonator pattern, expandable element) and (resonator pattern, flexible circuit assembly).

Table 7. Correlation of graphene sensing field fusion matrix (TOP 50).

No.	*Keyword1*	*Keyword2*	$M_{MB+MC+MP}$	*MB*	*MC*	*MP*
1	method	solution	0.361	1	0	0
2	method	patient	0.359	0.995	0	0
3	method	detecting	0.35	0.969	0	0
4	semiconductor	multi-walled carbon nanotubes	0.333	0	1	0
5	apparatus	discrete operative device	0.32	0	0.961	0
6	stimulation	discrete operative device	0.32	0	0.961	0
7	sensor	detecting	0.314	0.87	0	0
8	resonator pattern	expandable element	0.305	0	0	1
9	resonator pattern	flexible circuit assembly	0.305	0	0	1
10	layer	electrode	0.293	0.812	0	0
11	detecting	discrete operative device	0.285	0	0.854	0
12	reservoir	discrete operative device	0.285	0	0.854	0
13	patient	device	0.285	0.788	0	0
14	patient	system	0.28	0.775	0	0
15	layer	substrate	0.275	0.761	0	0
16	method	sensor	0.267	0.738	0	0
17	method	surface	0.266	0.736	0	0
18	hemorrhage	expandable element	0.259	0	0	0.848
19	hemorrhage	flexible circuit assembly	0.259	0	0	0.848
20	sensor	patient	0.255	0.706	0	0
21	method	glucose	0.254	0.703	0	0
22	nanowire	cadmium	0.252	0	0.756	0
23	method	chemistry	0.25	0.691	0	0
24	method	discrete operative device	0.249	0	0.747	0
25	cancer	discrete operative device	0.249	0	0.747	0
26	nerve	discrete operative device	0.249	0	0.747	0
27	parameters	discrete operative device	0.249	0	0.747	0
28	discrete operative device	electrical conductivity	0.249	0	0.747	0

29	discrete operative device	accurate detection	0.249	0	0.747	0
30	method	device	0.235	0.65	0	0
31	surface	detecting	0.234	0.649	0	0
32	method	electrode	0.234	0.647	0	0
33	layer	device	0.234	0.647	0	0
34	surface	device	0.233	0.645	0	0
35	resonator pattern	pressure sensor	0.232	0	0	0.759
36	method	graphene	0.231	0.638	0	0
37	electrodes	discrete operative device	0.23	0	0.689	0
38	method	sample	0.23	0.636	0	0
39	layer	sensor	0.227	0.627	0	0
40	substrate	lumen	0.22	0	0.66	0
41	surface	electrode	0.219	0.607	0	0
42	method	substrate	0.219	0.607	0	0
43	layer	discrete operative device	0.217	0	0.65	0
44	substrate	discrete operative device	0.217	0	0.65	0
45	device	detecting	0.217	0.6	0	0
46	resonator pattern	camera	0.216	0	0	0.708
47	sensor	substrate	0.216	0.597	0	0
48	material	expandable element	0.215	0	0	0.704
49	material	flexible circuit assembly	0.215	0	0	0.704
50	layer	surface	0.214	0.593	0	0

Comparison of three types of topic associations

To further study whether the three types of subject associations have complementary significance to each other, this study uses Jaccard similarity analysis (Jaccard, 1912) to compare the similarity between the three types of subject associations. The values of subject associations are divided into 0 and non-0 categories, and the similarities between the three relationships of terms co-occurrence, citation-terms coupling, and patent assignee-terms coupling are calculated respectively. The Jaccard distance is employed to quantify the similarity between the three sets of subject associations. It is noteworthy that the magnitude of this value directly corresponds to the extent of dissimilarity between the sets.

The repetition rate of non-0 elements in the three subject associations and the calculation results of Jaccard distance are shown in Table 8. The calculation results reveal that there is no overlap between the subject word co-occurrence association and the other two associations, and there is a very small amount of overlap between the citation-subject word coupling association and the patent applicant-subject word coupling association. The Jaccard distance between the three types of relationships is extremely high. It is evident that the subject features revealed by these three types of associations differ significantly and are highly complementary. Therefore, the integration of these three types of associations is conducive to enriching the clues of subject feature and enriching and deepening the significance of text mining.

Table 8. Similarity comparison of three types of topic association matrices in the field of graphene sensing.

Theme Relationship	*MB VS. MC*	*MB VS. MP*	*MC VS. MP*
Overlapping Relationship	0	0	278
Overlapping Rate	0%	0%	$\frac{\mathrm{MC} \cap \mathrm{MP}}{\mathrm{MC}} = 1.506\%$ $\frac{\mathrm{MC} \cap \mathrm{MP}}{\mathrm{MP}} = 13.385\%$
Jaccard Distance	1	1	0.986

The significance of multi-relationship integration

Based on the weighted fusion method of network credibility, this study obtained the fusion matrix of three types of associations. The variance and sparsity comparison of each matrix is shown in Table 9. The results show that the variance of the fusion matrix $M_{MB+MC+MP}$ is 0.013, which is smaller than the variance of the MB, MC, and MP matrix. At the same time, the sparsity of the fusion matrix is also notably lower than that of the three types of topic association matrix. The fusion matrix demonstrates a substantial degree of richness in information and exhibits a minimal degree of discreteness.

Table 9. Variance and sparsity of three types of topic association and fusion matrix in graphene sensing field.

Theme Relationship	*MB*	*MC*	*MP*	$M_{MB+MC+MP}$
Variance	0.029	0.049	0.092	0.013
Sparsity	0.836	0.953	0.995	0.784

Furthermore, as illustrated in Table 9, the MP matrix exhibits the highest degree of sparsity, indicating that the extracted relationship is relatively weak. The MC matrix demonstrates the second highest sparsity, while the MB matrix exhibits the lowest sparsity. Concurrently, as illustrated by Table 6, the fusion matrix allocates minimal importance to the MP matrix due to the limited number of feature relationships and the low level of information accuracy. A comparison of Tables 7 and 9 reveals the fusion matrix assigns a relatively low weight to infrequent relationships. However, it also attains 0.305. Therefore, infrequent relationships, despite their relative weakness, are represented to a certain extent in the fusion matrix. To elucidate this assertion, the association relationship of resonator pattern, expandable element and resonator pattern, flexible circuit assembly can be utilized as an instance. It is not expressed in the MB and MC matrices. However, it is observed that the fusion matrix (MP) assigns a weight of 1 to this relationship, signifying its significance. The weight value in the fusion matrix is 0.305. Within the 84,680 non-zero value relationships present within the fusion matrix, it is ranked 8th, indicating its significant contribution to the fusion matrix. This observation signifies that the fusion matrix does indeed consider weak relationships to a certain extent. It is evident that the

fusion matrix can not only effectively represent the primary attributes in multivariate relations but also possess a satisfactory degree of expressiveness for rare topic-related attributes.

Conclusions

This paper researches and proposes a multiple topic-association fusion method suitable for patent analysis. The proposed method is founded on the association features of multi-source topic data obtained in patent text mining. First, the paper extracts multiple topic correlations based on a number of multiple topic association relationships. Next, it combines a neural network and an evidence theory approach, creating a fusion method for multiple thematic correlations. The objective of this fusion method is to generate an information enhancement matrix that contains more comprehensive topic association relationships. Empirical study was conducted on the patent data in the domain of graphene sensing technology. It aimed to validate the efficacy of an integrated method of multiple thematic correlations. The method involves the learning of the weight distribution of different topic-associated relations through neural networks, the use of evidence theory to model and fuse the uncertainty of multi-source information, and the final generation of the topic-associated enhancement matrix. The empirical study demonstrated significant disparities among the three categories of subject association relationships: subject term co-occurrence, citation-subject term coupling, and patent applicant-subject term coupling. Their subject features manifested stronger complementarity. The fused matrix is characterized by enhanced informational content and reduced discreteness, resulting in a more comprehensive characterization of the primary subject association attributes. Additionally, the fused matrix is capable of expressing rare topic-weakly associated attributes with high efficacy. The integration of information from multiple sources through the fusion method, which is based on a neural network and evidence theory, has been shown to enhance the characterization of the association relationship between topics.

The following advantages and innovations of the proposed method are evident. Firstly, multi-source information fusion is achieved by combining neural networks and evidence theory, effectively combining three types of theme association relations: terms co-occurrence, citation-terms coupling, and patent assignee-terms coupling. This combination overcomes the limitations of a single association relation and significantly improves the comprehensiveness and accuracy of the theme association analysis. Secondly, uncertainty modelling is employed, which is another innovation. The modelling of uncertainty in multi-source information by evidence theory enhances the method's reliability and credibility. Furthermore, this enhanced uncertainty modelling provides a more robust foundation for analysing complex thematic association relationships. Enhancement of theme features: The enhanced theme association matrix, resulting from the fusion process, can effectively capture not only the primary theme attributes but also those of less prevalent theme weak associations. Consequently, this multifaceted approach provides a richer array of clues for the identification of technology themes and potential technological innovations.

Despite the findings of this study, there remain several issues that require further investigation. Specifically, there is a need for further refinement of the multiple topic association fusion method and its application. Optimization of computational efficiency is essential for addressing the challenges posed by the growth in size of the topic terms. This growth results in exponential expansion of the dimensionality of the matrix, necessitating high-performance computing capabilities. In future research, we will explore more efficient methods for dimensionality reduction processing of thematic feature terms (e.g., techniques based on graph embedding or sparse representation) to improve computational efficiency and reduce resource consumption. The verification of method universality is imperative. The present study principally focuses on the integration of three types of thematic associations, which can be further extended to more types of thematic associations (e.g., technological efficacy associations, technological evolution associations, etc.) in the future to verify the universality and robustness of the constructed method. The Dynamic Theme Modelling method will be employed to explore the evolving nature of the theme association relationships. The integration of dynamic theme modelling will facilitate the exploration of the temporal evolution of the theme association relationship, thereby providing a more profound foundation for the prediction of technological development trends and the prospective research of technological innovation. The method is applied to other technological fields (e.g., artificial intelligence, biomedicine, etc.) to verify its applicability and effectiveness in different fields and further expand the application scope of the method.

The multivariate topic association fusion method based on a neural network and evidence theory proposed in this study provides novel concepts and methodological support for patent text topic mining. As evidenced by the experimental findings, the proposed method has the capability of effectively integrating multi-source information, thereby enhancing the characterization of topic-association relationships. This, in turn, provides a powerful tool for technology topic analysis, technology trend prediction, and related fields. Subsequent research endeavours will focus on enhancing the method's computational efficiency, validating its universality, and incorporating dynamic topic modelling. These efforts aim to broaden the application of the method in diverse scenarios, thereby providing more intelligent and precise support for technological innovation and patent analysis.

Acknowledgments

The research is the outcome of the projects, “Youth Innovation Promotion Association (2022173)” “Light of West China program”, “Intelligence Service for Research Institutes on Science, Technology and Innovation (E3291106)”, “Research on Emerging Technology Direction Identification Method based on Technology Fusion (E3Z0000803)”, supported by Chinese Academy of Sciences(CAS), “Study on the Multi-Relation Data Fusion Methods for Identification and Prediction of Technology Innovation Paths”(No. 18BTQ067) supported by National Social Science Fund of China and “Early Recognition Method of Transformative Scientific and Technological Innovation Topics based on Weak Signal Temporal Network Evolution analysis” (No.72274113) supported by the National Natural Science

Foundation of China. Additionally, the special contribution of Xian Zhang as Corresponding Author (zhangx@clas.ac.cn) is noteworthy.

References

Chawla, S., Kaur, R., Aggarwal, P., 2023. Text classification framework for short text based on TFIDF-FastText. Multimed. Tools Appl. 82, 40167–40180. https://doi.org/10.1007/s11042-023-15211-5

Feng, G., Wu, J., Mo, X., 2020. Research on detection and verification of burst words with multiple measures. Library and Information Service 64, 67–76. https://doi.org/10.13266/j.issn.0252-3116.2020.11.008

Hamda, N.E.I., Hadjali, A., Lagha, M., 2023. Multisensor data fusion in IoT environments in dempster-shafer theory setting: An improved evidence distance-based approach. Sens. 23, 5141. https://doi.org/10.3390/s23115141

Jaccard, P., 1912. The distribution of the flora in the alpine zone.1. New Phytologist 11, 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Kleinberg, J., 2002. Bursty and hierarchical structure in streams, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02. Association for Computing Machinery, New York, NY, USA, pp. 91–101. https://doi.org/10.1145/775047.775061

Li, K., Dai, Z., Zuo, C., Wang, X., Cui, H., Song, H., Cui, M., 2024. Scene adaptation in adverse conditions: a multi-sensor fusion framework for roadside traffic perception. J. Intell. Transp. Syst. https://doi.org/10.1080/15472450.2024.2390844

Li, Z., Zhang, Q., Liu, S., Peng, Y., Li, L., 2024. Information fusion and attribute reduction for multi-source incomplete mixed data via conditional information entropy and D-S evidence theory. Appl. Soft Comput. 151, 111149. https://doi.org/10.1016/j.asoc.2023.111149

Lin, C., Kong, Y., Han, Q., Chen, K., Geng, Z., Wang, T., Dong, M., Liu, H., Chu, F., 2025. IF-EDAAN: an information fusion-enhanced domain adaptation attention network for unsupervised transfer fault diagnosis. Mech. Syst. Signal Process. 224, 112180. https://doi.org/10.1016/j.ymssp.2024.112180

Liu, Huijie, Wu, H., Zhang, L., Yu, R., Liu, Y., Liu, C., Li, M., Liu, Q., Chen, E., 2022. A hierarchical interactive multi-channel graph neural network for technological knowledge flow forecasting. Knowl. Inf. Syst. 64, 1723–1757. https://doi.org/10.1007/s10115-022-01697-2

Liu, Huailan, Zhang, R., Liu, Y., He, C., 2022. Unveiling evolutionary path of nanogenerator technology: a novel method based on sentence-BERT. Nanomaterials 12, 2018. https://doi.org/10.3390/nano12122018

Liu J., Zhong Y., He X., Li Z., Zhao Z., Wang H., Song S., 2024. Research on Scientific-Technological-Industrial Association Patterns Based on Multi-relationships Fusion. Journal of Modern Information 44, 67–81.

Pan, X., Wang, Y., He, S., 2021. The evidential reasoning approach for renewable energy resources evaluation under interval type-2 fuzzy uncertainty. Information Sciences 576, 432–453. https://doi.org/10.1016/j.ins.2021.06.091

Qaiser, S., Ali, R., 2018. Text mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications 181. https://doi.org/10.5120/ijca2018917395

Qian J., Tong Z., Yu Y., Hong C., Miao D., 2023. Multi-source information fusion through generalized adaptive multi-granulation. CAAI Transactions on Intelligent Systems 18, 173–185.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Song X., Chen M., Lyu G., Shen Y., 2023. A Research on Competitor Identification Model Based on Crossdomain and Multi-source Information Fusion in the Context of Big Data-Based on the New Energy Automobile Industry. Journal of the China Society for Scientific and Technical Information 42, 176–188.

Tan J., Yang A., Li T., Ma X., Chen S., 2022. Multi-source Fusion Navigation Algorithm Based on Robust Adaptive Filtering. Aerospace Control 40, 22–29. https://doi.org/10.16804/j.cnki.issn1006-3242.2022.05.012

Tseng, Y.-H., Lin, C.-J., Lin, Y.-I., 2007. Text mining techniques for patent analysis. Information Processing & Management, Patent Processing 43, 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

Xiao, F., 2023. GEJS: A generalized evidential divergence measure for multisource information fusion. IEEE Trans. Syst., Man, Cybern., Syst. 53, 2246–2258. https://doi.org/10.1109/TSMC.2022.3211498

Xu, H., Guo, T., Yue, Z., Ru, L., Fang, S., 2016. Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series. SCIENTOMETRICS 106, 583–601. https://doi.org/10.1007/s11192-015-1792-2

Xu W., Huang X., Cai K., 2023. Review of Multi-source Information Fusion Methods Based on Granular Computing. Journal of Data Acquisition and Processing 38, 245–261. https://doi.org/10.16337/j.1004-9037.2023.02.002

Yan J., Li H., Ma Y., Liu Z., Zhang D., Jiang Z., Duan Y., 2024. Multi-source Heterogeneous Data Fusion Technologies and Government Big Data Governance System. Computer Science 51, 1–14.

Yang, J., Liu, Y., Zhu, X., Liu, Z., Zhang, X., 2012. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Information Processing & Management 48, 741–754. https://doi.org/10.1016/j.ipm.2011.12.005

Yu, Y., Ju, P., Shang, M., 2022. Research on the evaluation method of patent keyword extraction algorithm based on information gain and similarity. Libr. Inf. Serv. 66, 108–117. https://doi.org/10.13266/j.issn.0252-3116.2022.06.012

Zhai D., Lou Y., Kan H., He X., Liang G., Ma Z., 2023. Constructing TCM Knowledge Graph with Multi-Source Heterogeneous Data. Data Analysis and Knowledge Discovery 7, 146–158.

Zhang, Q., Zhang, P., Li, T., 2025. Information fusion for large-scale multi-source data based on the dempster-shafer evidence theory. Inf. Fusion 115, 102754. https://doi.org/10.1016/j.inffus.2024.102754

Zhang, X., Lin, J., 2025. Scalable data fusion via a scale-based hierarchical framework: adapting to multi-source and multi-scale scenarios. Inf. Fusion 114, 102694. https://doi.org/10.1016/j.inffus.2024.102694

Zhang X., Zeng R., Li S., Li J., 2024. Patented Technology Evolution Path Identification Based on Multi-relation Data Fusion. Library and Information Service 68, 71–84. https://doi.org/10.13266/j.issn.0252-3116.2024.03.007

Zhang, Z., Wang, H., Geng, J., Deng, X., Jiang, W., 2024. A new data augmentation method based on mixup and dempster-shafer theory. IEEE Trans. Multimedia 26, 4998–5013. https://doi.org/10.1109/TMM.2023.3330106

Zhu, J., Deng, A., Xing, L., Li, O., 2024. Rolling bearing fault diagnosis based on multi-source information fusion. J. Fail. Anal. Prev. 24, 1470–1482. https://doi.org/10.1007/s11668-024-01935-5

https://doi.org/10.51408/issi2025_081

Research on Public Perception of Academic Achievements under Public Health Emergencies

Liu Xiaojuan[1], Hu Wei[2], Li Xinran[3], Xiao Yuntong[4]

[1]lxj_2007@bnu.edu.cn, [2]202321260034@mail.bnu.edu.cn, [3]202121260048@mail.bnu.edu.cn, [4]rsltong@163.com

School of Government, Beijing Normal University, No. 19, Xinjiekouwai Street, Haidian Beijing (China)

Abstract

The public perception of academic achievements under public health emergencies directly affects the recognition and release of the social value of the achievements. Analyzing this relationship will help improve the theories and methods of assessing the social impact of academic achievements. The study selected posts and user interaction data mentioning academic achievements on Weibo, a Chinese social media platform, during the COVID-19 pandemic as samples. Combining with public perception theory, we analyzed the public's comments and reposted texts, aiming to reveal the public's attention to academic achievements and their emotional attitudes. We found the public generally has a positive attitude of respect and trust toward academic achievements, researchers, and bloggers. The dissemination of academic achievements has a positive influence on the public's cognition and behavior. However, there are still some critical and questioning voices. In order to further improve the social impact assessment and promote the dissemination and influence of academic achievements among the public, it is recommended to fully explore the social media data that can be used for the social impact assessment, and build public trust in academic achievements through various stakeholders, such as researchers, mainstream media, and government departments.

Introduction

Since the 20th century, the interpenetration of science, technology, and society has gradually made scientific research a cause that requires the joint efforts of all sectors. Against the backdrop of the knowledge economy, scientific research, as the cornerstone for promoting national transformation and social progress, is being placed with greater social expectations. Compared with the academic impact, social impact driven by public values and social needs is gradually becoming an important consideration in the science and technology policies of many countries. In recent years, China has issued a number of policy documents emphasizing the assessment of the social impact of academic achievements. It has emphasized the

implementation of classified assessment and evaluation, focusing on the quality, contribution and impact of landmark achievements (Ministry of Science and Technology of the People's Republic of China, 2020); and pointed out that it was necessary to comprehensively and accurately assess the scientific, technological, economic, social and cultural value of scientific and technological achievements (The State Council of the People's Republic of China, 2021). Many international organizations have begun to conduct social impact assessments on a regular basis, such as the Research Excellence Framework (REF) in the UK, Research Quality Framework (RQF) in Australia, and Standard Evaluation Protocol (SEP) in the Netherlands.

As the knowledge economy continues to deepen and the model of knowledge production evolves, the public is no longer a passive recipient of knowledge (Fecher & Hebing, 2021). Scientific research assessment has increasingly focused on the social impact on the public. The dynamic four-spiral mechanism of Knowledge Production Model III, innovatively developed within the dual-spiral structure of the three-spiral nonlinear network model, has given rise to the "University-Industry-Government-Civil Society" innovation ecosystem model (Schütz, Heidingsfelder, & Schraudner, 2019), which affirms the important position of the public in scientific activities. In the era of self-media, the degree of engagement and activity in online science discussions has increased significantly. Several studies have demonstrated that social media platforms have significantly influenced research assessment by enhancing the visibility of scientific outputs, facilitating rapid dissemination, and promoting robust public engagement with research findings (Haustein, Costas, & Larivière, 2015; Sugimoto et al., 2017). Public participation in science not only improves their scientific literacy, but also influences the public cognition, values, and other aspects, thereby realizing the social value and broad dissemination of academic achievements. However, public attitudes toward science are often complex. On one hand, due to limited understanding of science, the public is willing to trust science and scientists as representatives of the scientific system, believing that science can solve problems. On the other hand, the uncertainties in science, negative events (such as academic misconduct), and the potential risks posed by scientific advances (such as genetically modified organisms and nuclear energy), often lead to public skepticism about scientists and scientific research. Furthermore, as science continues to develop toward sophistication, depth, and specialization, it becomes progressively more difficult for the public to fully understand science and technology. Consequently, the focus of relevant research has shifted from exploring whether

the public understands science to investigating whether the public trusts the science and scientists (Irzik & Kurtulmus, 2021; Goldenberg, 2023; Tranter, 2023).

Surveys show that global skepticism toward science has been on the rise (Nuyen, 2019), and the outbreak of COVID-19 in 2020 has exacerbated this challenge. The pandemic put science under the public microscope. Research issues are directly related to everyone's daily life, prompting the public to rely more on scientific research and expertise. During this period, mass media became a key source of scientific information for the public. The scientific community has also increasingly focused on communicating and interacting with the public through social media platforms, and social media data have been widely used in studies related to public trust. Van Dijck and Alinejad (2020) found that social media were indeed two-sided swords of health communication, and were deployed to both undermine and enhance public trust in scientific expertise during a health crisis. Algan et al. (2021) conducted a large-scale survey across twelve western countries from March to December 2020 and found a marked decline in public trust in scientists, particularly in France. Additionally, Mihelj, Kondor, and Štětka (2022) conducted a study involving interviews and diary surveys in four Eastern European countries, which revealed a general trust in experts. However, some respondents in Serbia and Hungary expressed strong distrust in the experts appointed to the national crisis teams by their governments. Public trust in science has been severely eroded by various sources of information, including paper retractions, the spread of pseudoscience on social media (Muhammed T, S., & Mathew, S. K., 2022), the spread of fake news triggered by flawed preprints, and research findings that fail to align with public expectations. Many people have started to question the professional competence, ethical conduct, and research motives of scientists, and these sentiments are spreading and being reinforced on social media. Positive or negative public perceptions of science have a direct impact on the public's acceptance and adoption of vaccines, therapeutic drugs, and public health policies based on scientific research, thereby affecting the ability of scientific research to achieve its societal value in improving health and well-being.

In summary, numerous studies have examined the relationship between the public and science, and the concept of social impact assessment of academic achievements is evolving toward focusing on stakeholders (Benneworth, 2017; Muhonen, Benneworth, & Olmos-Peñuela, 2020; Bonaccorsi, Chiarello, & Fantoni, 2021). Public attitudes toward science, especially on issues closely related to public interest, such as public health emergencies, play a crucial role in determining the real-world impact of scientific research and the stability of societal functioning. Therefore, it is

necessary to examine the public's views and attitudes toward academic achievements from the perspective of public perception. Users' activities on social media platforms, such as browsing, liking, commenting, and reposting, serve as primary means for users to express their opinions and engage in information exchange. These actions also reflect users' views, emotions, and cognition within the social media environment. Commenting and reposting, in particular, represent higher levels of user participation, as they involve more substantial cognitive and emotional engagement (Sailunaz & Alhajj, 2019). Therefore, this study, set against the backdrop of the COVID-19 pandemic, focuses on Weibo posts that mention academic achievements, along with their comments and reposts. Weibo, a mainstream social media platform in China, has 586 million monthly active users and high user engagement, making it an important channel for online communication and information gathering (Zhang, Jin, Liu, & Xue, 2024). By analyzing Weibo data, previous studies have offered crucial insights into the attitude and behavioral changes of Chinese social media users in the early stages of the COVID-19 pandemic(Li et al., 2020; Zheng, Adams, & Wang, 2024), and have also pointed out that the daily life status reflected by Weibo can help in predicting personality (Wang et al., 2020).By analyzing these comments and reposts, the study aims to explore the following questions:

Q1: How does the public perceive and understand academic achievements?

Q2: What attitudes does the public exhibit towards academic achievements?

Q3: What factors lead to the public's negative emotions and perceptions of academic achievements?

The findings will provide insights into enhancing public trust in science, promoting the social utility of academic achievements, and improving the assessment system for the social impact of scientific research.

Public perception theory

As key stakeholders in academic achievements, the public's attitudes and views directly influence the generation and dissemination of the social impact of academic achievements. Analyzing public perception is an effective way to understand these attitudes and views. The foundational theory of public perception suggests that public perception consists of cognition, emotion, and behavior, which together represent a social awareness of changes and effects in the objective world that impact one's own life (Qu & Lu, 2016). People judge unknown concepts or phenomena through cognitive processes, integrating them with their emotions or personal experiences. This leads to the formation of behavioral intentions, ultimately resulting

in consistent actions and perceptions. In this process, public perception plays a significant role in guiding group behavior. Related studies have defined public perception as the degree of awareness, attitudes, and views of the public toward specific events, issues, technologies, and policies (Stephanides, et al., 2019), or the knowledge and emotional attitude on specific topics (Huang et al., 2019； Fan & Zhuang, 2024). This study focuses on analyzing public perception of academic achievements through comments and reposts on Weibo. Drawing from the foundational theory of perception and previous research, we limit the scope of public perception to public attention and emotional responses to academic achievements. This study aims to provide deeper insights into the dissemination effects of scientific research in social media and how it influences public behavioral intentions. Ultimately, this research will help researchers and policymakers better understand and enhance the social acceptance of academic achievements.

Research design

Data collection

This study uses Weibo as the data source, focusing on popular accounts with high interaction and influence in the field of health and medicine. According to the survey, the top 10 most influential and the top 10 most popular influencers on Weibo in 2020 and 2021, as well as the most influential and most popular influencers in 2022, have been identified. Additionally, this study added the accounts of Nanshan Breathing (Nanshan Zhong's research team), Dr Zhang Wenhong, and five mainstream media accounts such as People's Daily, for a total of 48 source accounts. We used Python to scrape original posts containing the keywords "COVID-19", "novel coronavirus", "SARS-CoV-2" and "2019-nCoV" from these accounts. The time range for these posts is limited from January 2020 to June 2023. The posts mentioning academic achievements were classified into different subjects through manual categorization. The most prevalent subjects were "drug development", "epidemiological research", "virus structure, origin tracing, and infection mechanism studies". These subjects garnered the highest number of likes, comments, and reposts, indicating broad public interest and representativeness. Therefore, we selected posts within these three subjects as our sample.

We retained posts with more than 10 comments, and manually reviewed and filtered those that mentioned academic achievements according to the following criteria: (1) Posts mentioning papers, academic reports, vaccines, drugs, diagnostic technologies, and other types of scientific contributions are included in the dataset. (2) If the post

is related to COVID-19 but the mentioned academic achievement is not specifically relevant to the pandemic, that post is excluded. (3) Posts referencing academic achievements in various forms, such as links, images, references, DOIs, patent numbers, or those citing key elements like the journal, research team, or platform, or using key phrases like "research shows", "according to the literature", or "approved" are also included. After review and selection, the final dataset comprised 525 posts from the subjects of "drug development", "epidemiological research", "virus structure, origin tracing, and infection mechanism studies". Based on the unique ID of each post, we further crawled the first and second-level comments and repost texts, while gathering the number of likes on these comments and reposts to analyze public perception of academic outcomes.

Data coding

In posts mentioning academic outcomes, public comments, reposts, and other forms of interaction indicate the public engagement and the emotional responses to the research findings. The act of commenting itself demonstrates the public's interest, and the object of the comment further reveals their key concerns. Comment content often contains the public's specific opinions and feelings, serving as an external manifestation of public perception. Public perception of academic achievements shapes the direction and content of their comments. For instance, expressing personal opinions on the research outcomes and engaging in discussions and debates reflect the public's concern, while emotional reactions such as gratitude, praise, doubts, or criticism directed towards bloggers or researchers represent emotional feedback. Therefore, analyzing the content of the comments can further reveal the public's deeper perception of academic achievements.

Preliminary research indicates that public comments not only involve the academic achievements themselves, but also encompass various aspects, such as the credibility of researchers, bloggers' approaches to disseminating information, and the impact of related policies. Public attitudes toward researchers and bloggers may enhance or weaken public trust in research achievements, and criticism or questioning of related policies may also affect the practical application and public acceptance of academic achievements. Consequently, the public comments directed at various objects reflect the social impact of academic achievements from a multifaceted perspective. To this end, this study adopted the content analysis method and constructed a two-level coding system to categorize the comment content around various comment objects, including academic achievements, bloggers, researchers, and policies. This will help

grasp the different focuses in public discussions and fully understand the public's multi-dimensional perception characteristics of academic achievements.

In order to clarify the cognitive and emotional characteristics reflected in the content of users' comments on social media, this study reviewed related literature. Liu et al. (2017) identified three types of tweets quoting papers: excerpts from the paper, external information about the paper, and attitudes toward the paper. These attitudes can be further subdivided into positive, neutral, somewhat supportive and negative. Positive attitudes include not only emotionally positive tweets but also neutral or exploratory ones, such as speculation, humorous responses, linking the paper to other topics, raising questions, and potential applications of the findings.

Regarding speech acts, Searle (1976) first divided them into direct speech acts and indirect speech acts, and further divided the agent's behavior into elaboration or assertion, commitment, instruction, declaration, and expression based on basic conditions, sincerity conditions, prerequisites, and propositional conditions. Furthermore, Zhang et al.(2013) divided speech acts into statements, questions, suggestions, comments, and mixed categories. Nemer (2016) fully considered the characteristics of online communication and divided speech behaviors into asking, requesting, instructing, inviting, informing, claiming, expecting, accepting/rejecting, apologizing, thanking, etc. This study designed a comment coding scheme, as shown in Table 1, based on the comment motivation classification system from the relevant literature and preliminary analysis of the study's dataset. The primary category covers various objects of commentary, such as research outcomes, bloggers, researchers, policies, and others, all of which are assigned numeric codes. The secondary category focuses on the content of the comments, coded with lowercase letters.

Table 1. Comment coding scheme.

Objects of commentary	*Comments*	*Account for*
1-Academic achievements	a-Praise and recognition	Express praise, affirmation and recognition
	b-Criticism and questioning	Point out possible errors, question the scientificity and authenticity, etc.
	c-Discussion and conjecture	Discuss and propose conjectures
	d-Surprise and worry	Expressing negative emotions such as surprise and worry

Objects of commentary	*Comments*	*Account for*
	e-Recommendations and expectations	Propose suggestions for optimizing the achievements or future research directions, and express expectations for achievements
	f-Association	Share other achievements related to the results mentioned in the original blog, or the opinions of professionals
	g-Humour	Express opinions in a humorous and witty manner
	h-Statement of experience	Describe relevant experience or real situations based on achievements or blog content
	i-Mention of external information	Discusses external information such as publication journals, research teams, links, peer reviews, industrialisation status of achievements, etc.
	j-Communicating practical issues	Discuss real-life issues based on the achievements, such as precautions, how to apply it
2-Bloggers	a-Approval and thanks	Thank the blogger for sharing, agree with and support the blogger's point of view
	b-Criticism and questioning	Criticise or satirise the blogger's viewpoints and positions, question the correctness and objectivity of the post, point out errors in the content, etc.
	c-Suggestions	Suggestions to bloggers on how to improve the content of posts
	d-Asking questions	Consult bloggers about the problems existing in the practical application of academic achievements
3-Researchers	a-Praise and thanks	Express respect, trust and gratitude to researchers
	b-Criticism and questioning	Express doubt or sarcasm to researchers
4-Policy	a-Suggestions	Propose suggestions and expectations for policies
	b-Support and affirmation	Support or comply with policy arrangements

Objects of commentary	*Comments*	*Account for*
	c-Doubts and concerns	Express different opinions on policies and concerns about the impact of policies on personal lives
5-No clear object for comments	a- Incomprehension	Express difficulty in understanding the content of academic findings
	b-Praise	Directly express praise without naming the person or entity, and it is difficult to judge based on the original post
	c-Belief in science	Demonstrate belief in and support for science
	d-Firm beliefs	Express trust in China and good expectations for the future

Two coders randomly selected and pre-coded 10% of the comments from each subject. During the pre-coding process, we identified certain low-quality comments that either had little analytical value or were irrelevant to the research objectives of this study. These comments were excluded based on the following criteria: (1) posts with no substantive content, including reposts, @ other accounts, punctuation only, emojis, interjections, and so on; (2) comments containing profanity, inciting arguments, creating division, or engaging in personal attacks; (3) comments that only contained hashtags or replicated the content of the original post without contributing new insights; (4) incomplete or unclear statements; (5) comments unrelated to the content of the post, such as advertisements; (6) comments involving the politicization of science, such as conspiracy theories. The coders discussed their coding results to further clarify and refine the coding criteria. After excluding the above-mentioned types of comments, the consistency of the coding results exceeded 90%.

A single coder conducted formal coding, yielding a final dataset of 15,354 coded comments. This included 4,552 comments related to "epidemiological research", 3,213 regarding "viral structure, origin tracing, and infection mechanism studies", and 7,589 related to "drug development". A month later, a random 10% sample from each thematic category was selected for secondary coding, and the consistency coefficient exceeded 95%, demonstrating the reliability of this coding.

Public Perception of Academic achievements in Social Media

Focus and Attitude Analysis Based on Commentary Texts

Comments offer a direct means for users to express their opinions. These comment texts contain valuable original insights, from which we can extract the public's understanding, evaluation, and discussions of academic achievements. This helps to uncover specific public viewpoints. On the Weibo platform, many users express their support for a particular viewpoint by liking comments. As a result, comments with a high number of likes tend to reflect topics that attract widespread attention or recognition from the public, and have a high degree of dissemination and influence. This study focuses on the coding of comment texts related to posts that mention academic achievements, exploring the overall distribution of public comments and the content characteristics of comments with a high number of likes. The objective is to reveal the public's areas of focus and emotional attitudes toward academic achievements.

For Weibo posts mentioning academic achievements, the distribution of the objects of commentary is as follows: academic achievements (66.9%), bloggers (8.8%), researchers (4.8%), and policies (3.7%). In addition, comments without a specific object account for 15.9% of the total. This indicates that the public's primary interest lies in the academic achievements themselves, particularly their practical applications. In contrast, comments directed at bloggers, researchers, policies, or other aspects are less common. In this study, comments with 10 or more likes are defined as highly praised comments, and 1,441 comments were obtained from the screening, accounting for 9.4% of the total coded comments in the dataset. Figure 1 compares the content distribution of these highly praised comments to all comments.

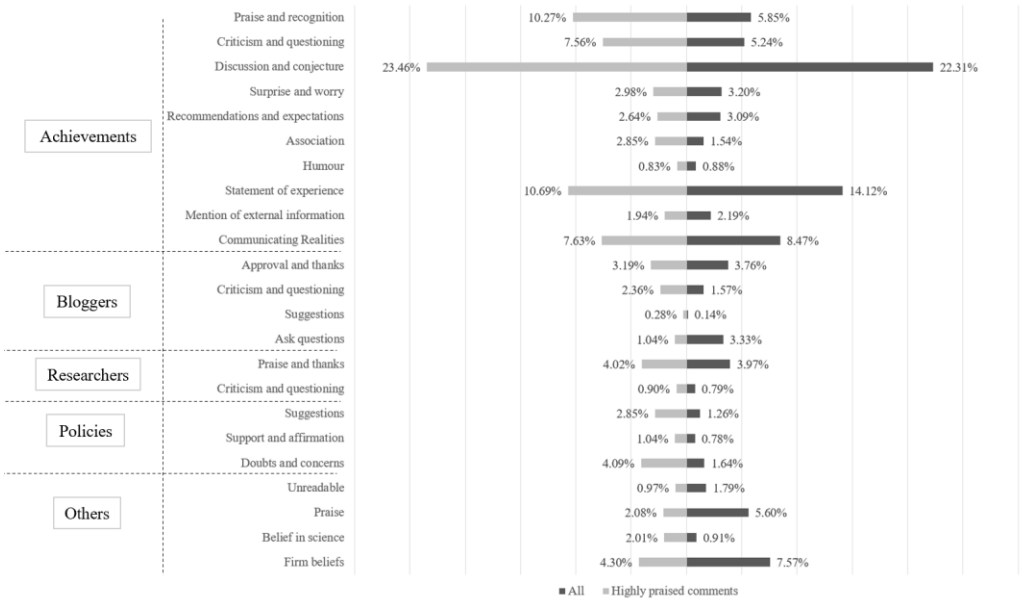


Figure 1. Comparison of content distribution of highly praised comments and all comments.

Focus and Attitude Analysis of Comments on academic achievements

Discussion and conjecture (22.3%) and statement of experience (14.1%) are the most common comment types on academic achievements. These are followed by communicating realities (8.5%), praise and recognition (5.8%), criticism and questioning (5.2%), surprise and worry (3.2%), recommendations and expectations (3.1%), mention of external information (2.2%), association (1.5%), and humor (0.9%). Weibo users typically engage in discussions about the details of academic achievements by combining the content of the original post, their own knowledge, and professional information sourced from other outlets. Despite the limited professionalism of public discussions, these interactions nonetheless demonstrate the significant public interest in academic achievements, a crucial aspect of their social impact. Simultaneously, the public also shows a greater concern for the practical application of these achievements in their daily lives, which is reflected in two content types: statements of experience and communicating realities. For instance, people might discuss the precautions that different groups should take when getting vaccinated or share their experiences after receiving the vaccine. In comparison to the overall percentage of comments, statements of experience and communicating realities receive significantly fewer likes. This is likely due to the clear association of such comments with individual attributes and specific needs, which limits their

widespread relevance. On the other hand, comments that express praise and recognition, as well as criticism and questioning, tend to receive more likes. This suggests that the public is more engaged with comments that clearly express a stance on the academic achievements. It also suggests that content with a clear and substantive attitude tendency can more accurately reflect the impact of achievements on public perception and has a higher analytical value.

Analyzing the focus and attitudes of comments directed at bloggers, researchers and policies

Among the comments on bloggers, approval and thanks (3.8%) and asking questions (3.3%) are the most common types of comments. The posts collected in this study came from Weibo-verified health bloggers and official mainstream media. These sources are widely recognized for their authority and professionalism, earning public appreciation. They also attracted inquiries on professional matters. This phenomenon demonstrates the public's trust in professionals and highlights their critical role in science communication on social media. Also, their involvement enhances the social impact of academic achievements. Among the comments directed at researchers, praise and thanks (4.0%) significantly outweigh criticism and questioning (0.8%). This suggests that the public's overall attitude towards researchers tends to be one of respect and trust. For comments directed at policy, doubts and concerns (1.6%) were the most frequent, followed by suggestions (1.3%). Support and affirmation (0.8%) were the least common. To a certain extent, this distribution shows the public's strong concern about policies based on academic achievements. However, these policies have not gained widespread recognition or acceptance. Policies play a crucial role in transforming and applying scientific research, directly affecting public life. For these policies to succeed, they must be adopted and followed by the public. Without public adoption, it will be difficult to achieve the intended outcomes, such as providing references for public policy formulation and safeguarding public health.

Focus and attitude analyses of comments without specified objects

Among the comments that did not specify the target audience, the most frequent were those expressing firm beliefs (7.6%), followed by praise (5.6%). These two categories even outnumbered the total number of comments directed at researchers and policies. Typically, these comments conveyed positive attitudes or firm beliefs in concise yet powerful language, often carrying strong emotional overtones. For example, expressions such as "fantastic", "go for it", "China will win", and "may the epidemic be overcome soon" appeared frequently. The large number of such

comments under posts mentioning academic achievements highlights the public's strong confidence in the power of scientific research to overcome the epidemic. However, compared with the overall percentage, comments expressing praise and firm beliefs are often brief and repetitive, which limits their ability to generate high engagement. Therefore, the percentage of highly praised comments is relatively lower. Moreover, these comments often lack substantive opinions about academic achievements and cannot clearly reflect the social impact of academic achievements. A small percentage (0.9%) of comments express belief in science, reflecting public faith in both the scientific community and its research achievements. On the other hand, 1.8% of comments indicate that the individuals "could not understand" the content, suggesting that a certain number of members of the public have difficulties understanding the content of the research achievements. This could undermine their trust in the academic achievements and thus be detrimental to the social impact of the research achievements. During the coding process, it was observed that when bloggers fail to appropriately simplify the original academic content, often directly quoting or translating it, the specialized language can become a barrier to public understanding. Some comments pointed out that "I usually just read the last two paragraphs of such articles because I can't understand the earlier parts" or "To be clear, we can't understand it because we're not studying medicine". This feedback suggests that communicators should emphasize the research findings most relevant to the public's daily lives and present them in simple, accessible language to improve understanding and acceptance. Taken together, the public's comments show a significant positive trend, reflecting the positive social impact of the research achievements on the public.

Analysis of attitudinal tendencies based on reposted texts

Users' reposting behavior on Weibo reflects the process of information diffusion and their selective attention to specific content, which highlights the public's recognition of the information. The analysis of the reposted text could reveal the dissemination and potential influence of the academic achievements in social networks. After conducting word frequency analysis on the coded comment texts, this study identified terms that can help specify the comment objects, including bloggers, researchers, and academic achievements. By using the co-word analysis in the reposted texts, it is possible to further judge the public attitude toward various objects. The results can be used for social impact analysis.

Attitudinal tendencies toward bloggers

The word frequency statistics of the comment texts targeting bloggers reveal high-frequency words that indicate the object of comments, including "blogger", "teacher", "editor", and "doctor". From the processed reposted texts, we extracted 422 entries containing these terms, representing 1.5% of the total data. Figure 2 displays the word cloud of the top 100 words that co-occur with the above words, with word size reflecting the number of co-occurrences. According to the word cloud, the overall attitude of the public toward bloggers in the reposted texts shows a tendency of trust and gratitude. High-frequency co-occurring words such as "Science popularization", "believe", "thank you", "need", "professional" and "hope" reflect the public's high recognition and appreciation for bloggers' science popularization activities through Weibo. These professional interpretations enhance the public's trust in academic achievements and deepen their understanding of them. Additionally, the public initiates interactions with bloggers using phrases such as 'Hello,' 'May I ask' and 'Could I inquire', which correspond to the 'Asking questions' comment type, further highlighting the trust in bloggers' expertise.

Attitudinal tendencies toward researchers

The word frequency statistics of the comment texts targeting scientific researchers reveal that the high-frequency words that can specify the object of the comments include "scientific researchers", "scientists", "professionals" and "researchers". We extracted 420 data points containing the above words from the processed reposts, representing 1.5% of the total number of reposts. Fig. 3 shows the word cloud of the top 100 terms that co-occur with these words, indicating researchers. According to the word cloud, the words that co-occur more often with researchers include "hard work", "China", "keep going", "gratitude", "respect" and "thank you" etc. These words reflect the public's recognition, gratitude and respect for researchers. At the same time, words such as "impressive", "hope", "great", "success", "believe", "effort" and other positive words are also displayed, further highlighting the public's positive attitude towards researchers. It can be observed that the reposted texts, similar to the comment texts, show an overall positive attitude of the public towards researchers.



Figure 2. High-frequency co-occurrence word cloud of reposted texts toward bloggers.



Figure 3. High-frequency co-occurrence word cloud of reposted texts toward researchers.

Attitudinal tendencies toward academic achievements

We selected five high-frequency words that can represent academic achievements and counted the adjectives that co-occur more frequently with these words, as shown in Table 2. Based on the statistical results, these high-frequency co-occurring adjectives are predominantly positive in sentiment, although a small number of negative emotion words are also present. These adjectives offer more reference value for the assessment of research achievements in the field of health and medicine. Positive words such as "significant", "effective" and "take effect" indicate that the public is positive about the efficacy of drugs. In addition, some universal positive adjectives were also mentioned frequently, including "awesome", "powerful", "successful", "best", "important", etc. These words not only reflect the public's praise and trust in research achievements but also imply that these achievements have a positive impact on public perception. The public is willing to actively disseminate these valuable academic achievements, which helps promote their acceptance and application, such as increasing the public's willingness to vaccinate.

Table 2. High frequency co-occurrence adjective list for academic achievements.

Representative terms of academic achievements	*High-frequency co-occurring adjectives (number of co-occurrences)*
Research	Important (34), Effective (29), Significant (19), Popular (13), Best (12), Obvious (9), Awesome (9), Unique (8), Reliable (7), Strict (7)
Miracle drugs	Effective (83), Awesome (6), Successful (5), Powerful (3), Important (3), Powerful (3), Great (3), Failed (2), Unnecessary (2)
Traditional Chinese medicine	Effective (53), Powerful (36), Awesome (33), Profound (33), Mighty (21), Great (11), Dependable (8), Useful (7), Fantastic (7), Proud (7)
Vaccinations	Effective (108), Urgent (63), Successful (41), Ineffective (30), Significant (27), Efficient (18), Serious (18), Best (18), Important (14), Powerful (13)
Data	Good (15), Effective (13), Best (12), Very good (8), Reliable (6), Strict (6), Important (6), Cautious (5), Obvious (5)

Analysis of the causes of negative perceptions based on object of comment

Based on previous analyses of the coded comments, this study found that the public's perceptions of posts that mention scientific achievements exhibit a range of emotional tendencies, including positive, negative, and neutral. According to trust theory, trust is a "leap of faith" or willingness to be vulnerable (Leith, 2013). The public needs scientific knowledge to solve problems and dilemmas in practice and tends to trust science and researchers with expertise, seeking their assistance. However, this trust can be altered by a variety of factors, which in turn can affect the public's acceptance and support of specific scientific achievements. It is clear that only when researchers and their achievements are trusted and accepted by the public can achievements be integrated into public practice and have a positive social impact. In contrast, negative perceptions, such as public criticism and questioning, can impede the realization of social impact. In order to promote the social benefits of research achievements, this study analyzes qualitatively the comment texts around bloggers, researchers, and academic achievements, especially those expressing negative attitudes, and analyzes what causes the public's negative perception of academic achievements.

Negative perception analysis toward bloggers

Based on the bloggers' identity authentication on Weibo, this study classified the bloggers into four identity types: mainstream media, clinician, R&D personnel, and industry association members. By extracting the comments directed at the bloggers, the distribution of comment content was plotted as shown in Figure 4. Criticism and questioning, which express negative public attitudes, accounted for the highest proportion of comments in mainstream media. However, comments directed at

clinicians, R&D personnel, and industry association members mainly consisted of approval, gratitude, and questions.

Tracing back to the original text, the study found that the main reason for the public's negative perception of academic achievements was the questioning of the professionalism and authenticity of the posts. This is manifested in the following ways: the content published by bloggers exhibits problems of lack of rigor, such as inconsistent images and typos, as well as unprofessional issues like misinterpretation of research conclusions and the incorrect use of professional terms. In addition, the completeness and objectivity of the published content also affect the public's perception of academic achievements. For example, some bloggers either failed to provide accurate data on vaccine protection rates, lacked detailed explanations on sample selection, or generalized conclusions based on limited samples. As a result, this may lead to doubts or misunderstandings among the public about the findings of scientific research and thus may hinder the positive impact of academic achievements on the public's perceptions and behaviors. Meanwhile, in the process of research dissemination, bloggers often neglect to explain the research design and focus only on presenting the research findings, and there are also a number of inaccurate citations or ambiguous statements. To ensure that academic achievements have a positive impact on public perceptions and are effectively applied, communicators need to be rigorous and precise in their references to the achievements, elaborating on key points that the public may have doubts about.

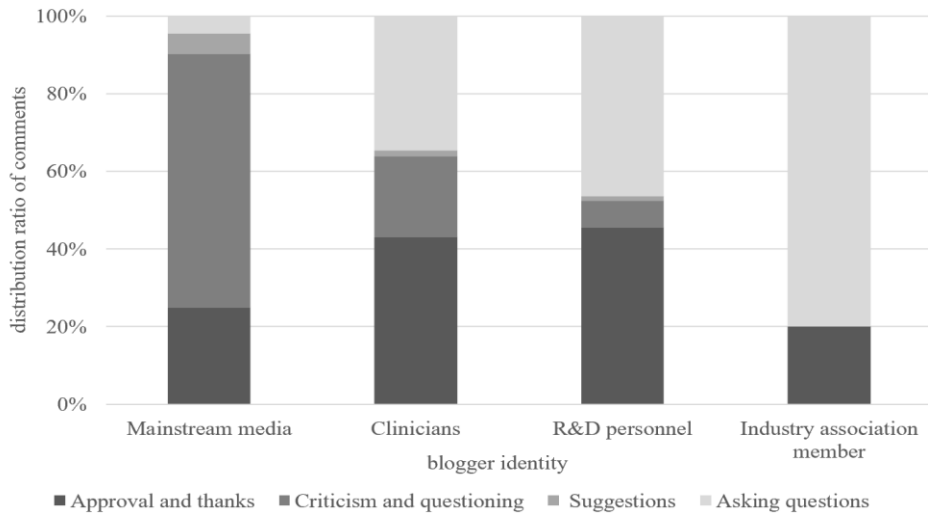


Figure 4. Distribution of comments by blogger identity.

Analysis of negative perceptions toward researchers

According to the coding scheme, the content of criticism and questioning reflects the distrust of researchers, with a total of 121 comments. Due to the relatively small sample size, this study identified the following three key reasons contributing to the public's negative perception of researchers through manual reading and analysis. Firstly, the public believes that researchers fail to prioritize the public's interest in conducting scientific research. Instead, they are perceived as being more driven by personal reputation and professional appraisal, thus making it difficult to understand and respond to the public's actual needs and plight. Secondly, the public lacks confidence in the research process, feeling uncertain about the reliability and validity of researchers' work. Doubts about the rigor of research methods and perceived inadequacies in research practices directly influence the public's trust in the research outcomes. Thirdly, the public points out that researchers' remarks in public lack objectivity and fail to reflect their professionalism and rigor. This undermines public trust in researchers.

In summary, the public's negative perception of researchers primarily arises from two phases: research process and dissemination of results. This distrust may further lead to negative public perceptions of research achievements, hindering their social application and overall impact. To address these concerns, it is recommended that researchers pay more attention to the public value of their research work when conducting research, and ensure that research projects can effectively respond to public concerns. During the research process, researchers should strictly adhere to research ethics and academic norms, ensuring the transparency and scientific integrity of the research to enhance its rigor and credibility. When disseminating research achievements, researchers should maintain an objective and professional attitude, clearly and accurately presenting the findings to foster a positive public image and facilitate the effective dissemination and application of research outcomes.

Analysis of negative perceptions toward academic achievements

According to the coding scheme, the criticism and questioning content reflects public mistrust toward academic achievements, with a total of 804 comments. This study used the LDA model to classify the public's negative perceptions toward the achievements into three themes. Figure 5 presents the co-occurrence map of theme words. Themes are distinguished by different grayscale levels in the figure, as shown in the legend.

Terms like "data", "vaccine", "study", "virus", "sample", "mortality", "drug", "effect" and "conclusion" were found across multiple themes, indicating that these terms are central to public concerns. Theme 1 includes terms such as "experiment", "trial", "clinical trial", "double-blind", "side effect", "placebo", "control group", "traditional Chinese medicine", "chloroquine" and "hydroxychloroquine". These terms reflect the public's distrust of the design of clinical trials of drug effectiveness and their implementation. The public may be concerned about the scientific validity of the trial design, the reliability of the trial results, and the risk of potential side effects. Theme 2 includes terms like "patient", "sample size", "statistics", "proportions" and "blood type". These terms point to the public's questioning of sample selection in research achievements. The public believes that sample selection bias may affect the accuracy and representativeness of the research findings, which may lead to discrepancies between the findings and the actual situation. Theme 3 includes terms such as "R&D", "strain", "number", "mutation", "efficacy", "infection rate", "protection rate" and "test". These terms reflect public doubts about the protective effects of vaccines. The public may have concerns about the vaccine development process, its effectiveness against different strains, and its overall efficacy.

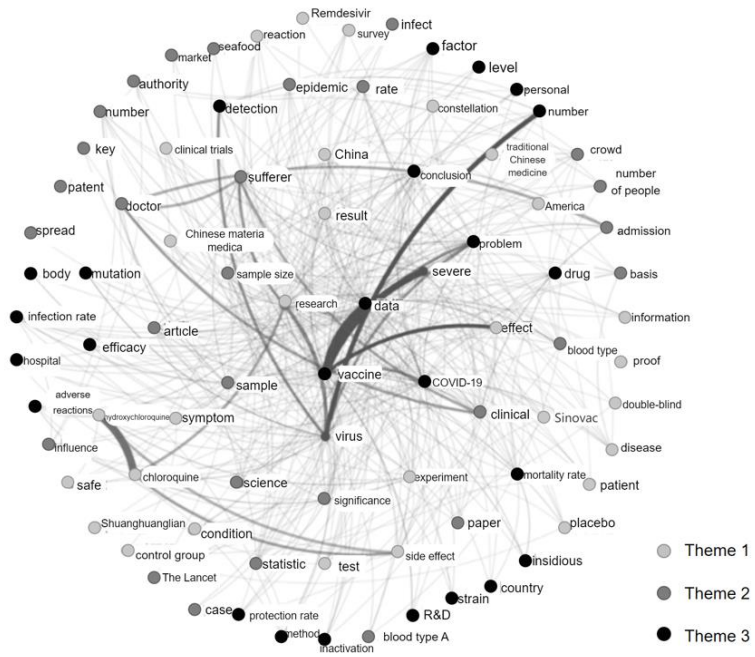


Figure 5. Thematic Co-occurrence of Negative Perceptions of Academic Achievements.

Conclusion and Implications

In this study, we used Weibo as the data source, collecting posts, comments, and reposts that mentioned academic achievements related to the COVID-19 pandemic, from health-related bloggers and mainstream media. Combining with the theory of public perception, we designed a coding scheme for comments and manually coded them. The public attitudes and perceptions reflected in likes, comments, and reposts were used to analyze the social impacts of academic achievements. The study shows that the public generally holds a positive attitude of respect and trust toward academic achievements, researchers, and bloggers. The dissemination of scientific findings has had a positive influence on public cognition and behavior, though some critical and skeptical voices still remain. In order to further improve the social impact assessment of academic achievements and enhance their dissemination and influence among the public, this study puts forward the following suggestions:

Fully explore social media data for assessing the social impact of academic achievements

With the development of the Internet era, an increasing number of the public access the latest scientific information and participate in public discussions through online media. The content of academic achievements and their applications (e.g., policies, products), as well as the corresponding user comments and reposts data, can be used as an important source of data for assessing the social impact of academic achievements. Based on the coding results of comments in this study, the content of public comments is varied and complex. On one hand, there are many comments that are unrelated to academic achievements or lack substantial content. These comments can hardly reflect the actual social impact of the research achievements. Therefore, the corresponding machine learning algorithms such as similarity matching, keyword recognition need to be developed to filter and mine online texts for social impact analysis. On the other hand, among the valid comments related to academic achievements, the focus of the commenters varies, such as academic achievements, bloggers, researchers, policies, etc. These objects of commentary are directly or indirectly linked to the impact of research achievements and can reflect the social impact from different perspectives. We should assign different weights based on the content of the comments when conducting social impact assessments. For example, regarding comments toward bloggers, we should consider the blogger's attitude toward the academic achievements in the original post to judge whether the social feedback of the achievements is positive or negative. Comments toward policies should be given higher weight, as they directly reflect the practical application of the

research outcomes. Comments that validate the research conclusions with personal experience, although showing public support for science, should be assigned less weight because they lack a professional perspective and are highly subjective. Similarly, comments that merely repeat the content of the research outcomes without offering new insights should be assigned lower weight.

Enhancing Public Trust in Academic achievements from the Perspective of Multiple Stakeholders

Enhancing public trust in academic achievements can promote the public's acceptance and application of scientific research results and secondary outputs based on them. This is of enormous significance for promoting the full use of academic achievements. To this end, public trust can be enhanced by focusing on the key stakeholders involved in the social impact transmission mechanism of research achievements: researchers, mainstream media, and government departments.

Researchers are both producers of academic achievements and the main force of scientific communication. Public mistrust of researchers primarily emerges during the phases of conducting research and disseminating research outcomes. Therefore, in fields closely related to public interests, scientific researchers should enhance their social responsibility, designing research topics and conducting studies based on public needs and interests. Scientific researchers should maintain a rigorous attitude toward their research work and avoid engaging in academic misconduct. Furthermore, researchers should actively engage in science communication by using social media platforms to share and exchange scientific information with the public, thereby bridging the gap between the public and academia. Our study has shown that the public tends to trust clinical doctors and researchers more than mainstream media, indicating that researchers' involvement in science communication activities can better enhance public understanding and acceptance of research achievements. However, as holders of specialized knowledge, researchers must align their communication with the actual needs of the public, providing clear answers to the scientific questions that the public cares about and simplifying technical language, explaining terms when necessary, rather than directly copying research texts. Our study found that a significant number of public comments expressed confusion, such as, "I don’t understand." Related research has also shown that different readability characteristics affect the Altmetric Attention Score of academic papers (Jin et al., 2021). Moreover, public skepticism toward research outcomes is partly due to insufficient explanations of research content. So, one of the challenges for scientists in science communication is ensuring that complex research processes and

conclusions are explained in a clear, simple, and objective manner. This is key to whether the public will truly recognize and accept research achievements, ultimately generating the desired social impact.

Mainstream media, with their broad audience reach and significant influence, have become the primary channels for the public to obtain scientific information, carrying the important responsibility of guiding public opinion. According to this study's analysis of how mainstream media mentions research achievements, media outlets tend to present the latest research findings succinctly, focusing on disseminating and promoting outstanding research achievements. Their reports are typically short, concise, and to the point. Although the main task of the mainstream media is not to analyze academic achievements in depth or answer public questions, the study found that they have used misspelled words, misused professional terms, and misinterpreted the research conclusions. These unprofessional actions can mislead the public to some extent and damage public trust in science. Therefore, while striving for timeliness in news reporting, mainstream media should maintain a rigorous and objective attitude. They must carefully verify information sources and present scientific information accurately and in detail, avoiding sensationalism, exaggeration of research findings, and improper inferences about research outcomes. To this end, mainstream media could establish a collaborative mechanism with researchers or professional science communicators to review content professionally before publishing related news reports, ensuring the authenticity and accuracy of the information.

Government departments develop public policies based on scientific research findings to promote the enhancement of public health and well-being, as well as the advancement of socio-economic development, thereby allowing research achievements to achieve their ultimate social impact. The scientific validity and rationality of public policies, as well as their ability to reflect the fundamental interests of the public, directly influence how the public understands and implements these policies. This study found that the public has voiced negative sentiments regarding certain policies on social media, indicating potential concerns or misunderstandings about the policy content or formulation process. To enhance public recognition and compliance, government departments should ensure transparency in using scientific research to formulate policies. This includes clearly stating the theoretical basis, scientific principles, and expected social effects of the policy, ensuring that the policymaking process respects and reflects the public's fundamental interests. Furthermore, government departments should establish comprehensive communication mechanisms to explain the background, objectives,

anticipated outcomes, and potential challenges of policies through diverse channels. This will help enhance public understanding and foster trust and support for the policies.

This study has some limitations. Firstly, due to the selection of Weibo as the data source, its built-in information filtering and blocking mechanisms resulted in the inability to access some negative comments. Secondly, the study limited the sources of posts to mainstream media and high-influence health domain bloggers, making the field of study somewhat narrow. To address these limitations, future research will expand the data sources, considering the inclusion of data from platforms such as Zhihu and WeChat official accounts. Additionally, future research will cover a broader range of research fields, focusing on research outcomes with high public attention and those closely related to public interests.

References

Ministry of Science and Technology. (2020). Measures to eliminate the negative bias of "paper-only" evaluation in scientific research (trial implementation). Retrieved February 23, 2020 from: https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2020/202002/t20200223_151781.html.

General Office of the State Council. (2021). Guiding opinions on improving the evaluation mechanism for scientific and technological achievements. Retrieved August 2, 2021 from: https://www.gov.cn/zhengce/zhengceku/2021-08/02/content_5628987.htm

Fecher, B., & Hebing, M. (2021). How do researchers approach societal impact? PLoS One, 16(7), e0254006.

Schütz, F., Heidingsfelder, M. L., & Schraudner, M. (2019). Co-shaping the future in quadruple helix innovation systems: Uncovering public preferences toward participatory research and innovation. She Ji: The Journal of Design, Economics, and Innovation, 5(2), 128-146.

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. PloS One, 10(3), e0120495.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. Journal of the Association for Information Science and Technology, 68(9), 2037-2062.

Irzik, G., & Kurtulmus, F. (2021). Well-ordered science and public trust in science. Synthese, 198, 4731-4748.

Goldenberg, M. J. (2023). Public trust in science. Interdisciplinary Science Reviews, 48(2), 366-378.

Tranter, B. (2023). Do Australians trust scientists? It depends on the 'science'. Australian Journal of Social Issues, 58(4), 821-837.

Nuyen, S. (2019). People are growing more skeptical of science, study finds. Retrieved March 20, 2019 from: https://www.wbir.com/article/news/nation-world/people-are-growing-more-skeptical-of-science-study-finds/507-10a134fb-51dd-40d9-ad74-19e2f246e71e.

Van Dijck, J., & Alinejad, D. (2020). Social media and trust in scientific expertise: Debating the Covid-19 pandemic in the Netherlands. Social Media + Society, 6(4), 2056305120981057.

Algan, Y., Cohen, D., Davoine, E., Foucault, M., & Stantcheva, S. (2021). Trust in scientists in times of pandemic: Panel evidence from 12 countries. Proceedings of the National Academy of Sciences, 118(40), e2108576118.

Mihelj, S., Kondor, K., & Štětka, V. (2022). Establishing trust in experts during a crisis: Expert trustworthiness and media use during the COVID-19 pandemic. Science Communication, 44(3), 292-319.

Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: a review of research in social media. International journal of data science and analytics, 13(4), 271-285.

Benneworth, P. (2017). We need better understanding of'good'research impacts.

Muhonen, R., Benneworth, P., & Olmos-Peñuela, J. (2020). From productive interactions to impact pathways: Understanding the key dimensions in developing SSH research societal impact. Research Evaluation, 29(1), 34-47.

Bonaccorsi, A., Chiarello, F., & Fantoni, G. (2021). Impact for whom? Mapping the users of public research with lexicon-based text mining. Scientometrics, 126(3), 1745-1774.

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. Journal of Computational Science, 36, 101003.

Zhang, J., Jin, G., Liu, Y., & Xue, X. (2024). Attention and sentiment of Chinese public toward rural landscape based on Sina Weibo. Scientific Reports, 14(1), 13724.

Li, J., Xu, Q., Cuomo, R., Purushothaman, V., & Mackey, T. (2020). Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance study. JMIR Public Health and Surveillance, 6(2), e18700.

Zheng, P., Adams, P. C., & Wang, J. (2024). Shifting moods on Sina Weibo: The first 12 weeks of COVID-19 in Wuhan. New Media & Society, 26(1), 346-367.

Wang, P., Yan, Y., Si, Y., Zhu, G., Zhan, X., Wang, J., & Pan, R. (2020). Classification of proactive personality: Text mining based on weibo text and short-answer questions text. Ieee Access, 8, 97370-97382.

Qu, Z., & Lu, Y. (2016). Finding the essence and iterative logic of public perception. Studies in Dialectics of Nature, 32(4), 96-101.

Stephanides, P., Chalvatzis, K. J., Li, X., et al. (2019). Public perception of sustainable energy innovation: A case study from Tios, Greece. Energy Procedia, 159, 249-254.

Huang, L., Wang, X., Wu, F., et al. (2019). Research on public perception of innovation policy based on network information mining: A case study on new energy vehicle policy. Science of Science and Management of S. & T., 40(6), 21-36.

Fan, H., & Zhuang, Y. (2024). Study of public perception of the metaverse based on content mining on Zhihu platform. Journal of Modern Information, 44(2), 65-80.

Liu, X. Z., & Fang, H. (2017). What we can learn from tweets linking to research papers. Scientometrics, 111(1), 349-369.

Searle, J. R. (1976). A Classification of Illocutionary Acts. *Language in Society*, *5*(1), 1–23.

Zhang, R., Li, W., Gao, D., & Ouyang, Y. (2013). Automatic Twitter topic summarization with speech acts. IEEE Transactions on Audio, Speech, and Language Processing, 21(3), 649-658.

Nemer, D. (2016). Celebrities acting up: A speech act analysis in tweets of famous people. Social Networking, 5(1), 1–10.

Dong, T., Xia, H. L., & Liang, C. (2013). An empirical study on the characteristics of People's Daily's official microblog news commentary based on "Hello, Tomorrow". Contemporary Communication, 40(4), 14-18.

Gao, J. (2016). Analysis of the participation motivation of information audience body in the network opinion matters. Library and Information Service, 60(9), 91-98.

Leith, D. (2013). Representations of the concept of trust in the literature of library and information studies. Cosmopolitan Civil Societies: An Interdisciplinary Journal, 5(3), 54–74.

Jin, T., Duan, H., Lu, X., et al. (2021). Do research articles with more readable abstracts receive higher online attention? Evidence from Science. Scientometrics, 126(10), 8471–8490.

https://doi.org/10.51408/issi2025_072

Research on the Innovation Performance Improvement Path of Scientific Research Team from the Perspective of Network Embeddedness

Junwan Liu[1], Qiqi Zhang[2], Shuo Xu[3], Chenchen Huang[4], Xiaoyun Gong[5], Jiahao Li[6]

[1]*liujunwan@bjut.edu.cn*, [2]*btbuzqq@163.com*, [3]*xushuo@bjut.edu.cn*, [4]*huangchenchen@bjut.edu.cn*, [5]*13718287848@163.com*, [6]*lijh0707@emails.bjut.edu.cn*

College of Economics and Management, Beijing University of Technology, Beijing (China)

Abstract

The scientific research team, as a crucial element of scientific and technological innovation, relies on the optimization of a collaborative network to expedite the enhancement of team innovation performance. Thus, it is of great significance to understand the factors affecting teamwork performance from the perspective of collaboration network. From the network embeddedness perspective, this study take scientific teams, technological teams and science-technology teams identified from literature and patent data in the pharmaceutical field from 1993 to 2019 as research sample, and applies the CART decision tree algorithm to explore the interactive effects on the innovation performance of different types of research teams and the improvement paths in terms of the three major network characteristic dimensions of structural embeddedness, relational embeddedness, and content embeddedness. The results show that: (1) Innovation performance improvement is not a simple function of single variables but a complex interaction of multidimensional factors. Different combinations of central position, intermediary position, collaborative tie strength, academic age diversity and knowledge interdependence will lead to the improvement of team performance. Knowledge interdependence is a key factor influencing innovation performance across all types of research teams. (2) In scientific teams, low academic age diversity and fewer intermediary positions foster close collaboration, enhancing novelty innovation performance. Teams with high academic age diversity and low knowledge interdependence are more likely to achieve breakthrough innovations by challenging path dependence. For impact, members with high knowledge interdependence and central positions effectively leverage accumulated knowledge and influential connections, significantly boosting the team's innovation performance. (3) In technological teams, low knowledge interdependence enables breakthroughs in technological barriers, fostering innovation and novelty. Members with high knowledge interdependence but low central positions benefit from independent R&D, mitigating network homogenization. For impact, teams with diverse academic ages and high knowledge interdependence leverage their knowledge tradition and varied perspectives to navigate technological iterations, forming unique technological systems and achieving significant impact. (4) In science-technology teams, academic inventors with low central positions have greater autonomy, enabling flexible resource allocation and deeper exploration of scientific frontiers. High centrality combined with strong collaborative ties fosters pioneering innovations through shared knowledge and norms. When collaborative tie strength is low, academic age diversity can compensate, driving high-performance novelty. For impact, low collaborative tie strength and diverse academic ages help minimize conflicts while leveraging varied researcher resources, continuously strengthening the team's influence.

Introduction

Teams are the engines of modern science, having grown in both prevalence and size across all areas of scientific and scholarly investigation and have become a prevalent research pattern in science and technology (Fortunato et al., 2018; Hao et al., 2024). Innovation is increasingly vital for teams seeking new opportunities, improved performance, and competitive advantage (Harvey & Berry, 2023). Network embeddedness fosters an environment in which innovative entities can acquire, apply, share knowledge, which empowers research teams to break through pyramidal innovation performance restrictions, and swiftly access valuable information resources to tackle complex scientific questions (Tian et al., 2021; Salazar & Lant, 2018). However, network embeddedness has a variety of mechanisms to drive innovation performance, which requires different resource bases and organizational modes, and may trigger synergistic or competitive relationships (Dong et al., 2018). Therefore, a thorough comprehension of the correlation between network embeddedness and innovation performance is crucial for the effective formation and optimization of teams, as well as for facilitating significant advancements in innovation.

Several studies have been conducted on the effect of scientific collaboration network embeddedness, holding a view that it is an essential factor that impacts the success of research teams (Zaheer et al. 2010). Network embeddedness is typically discussed in terms of three dimensions: structural embeddedness, relational embeddedness, and content embeddedness (Rishika et al., 2019; Mark, 1992). Structural embeddedness refers to the differences in the network positions within team collaboration network, including central and intermediary positions of research teams, which can indirectly provide teams with expanded information resources, thereby enhancing their capacity for innovation (Yan et al., 2019; Kramolis and Svirakova 2019). However, as network embeddedness continues to improve, team cohesion is strengthened. Once it reaches a certain level, the connection between the team and external innovation entities may gradually weaken, potentially leading to exclusivity and forming a local lock, thus hindering the reception of new knowledge and affecting the stimulation of creativity (Salazar & Lant, 2018). Regarding relational embeddedness, some studies have examined the impact of network strength, confirmed that the relational embeddedness and breakthrough innovation performance are inverse-U related (Deng et al., 2023). Content embeddedness refers to the various qualities that team members are expected to possess in their academic or professional endeavors, and typically includes academic age diversity, research topic diversity, and knowledge interdependence, which can enhance the cognitive flexibility of innovation teams, thus providing broader solutions to innovation challenges (Zhou et al., 2024). However, it may also give rise to conflicts and disconnection phenomena within the team (Haeussler & Sauermann, 2020). Therefore, although many scholars have explored the influence of network embeddedness on innovation performance, the results remain controversial. Meanwhile, these studies typically emphasize the independent impact of a single factor on team performance, frequently overlooking the impact of different configurations of multidimensional team features on high-level innovation. In other

words, it is yet to be determined how specific combinations of diverse team features affect the likelihood of achieving breakthroughs in science and technology remains uncertain (Lyu et al., 2021).

Additionally, the difference in factors affecting the network innovation performance of scientific, technological, and science-technology teams (collectively referred to as research teams) have been overlooked. Scientific teams typically consist of paper authors who are closely interconnected within co-authorship networks. By promoting knowledge integration among members, they drive the development and dissemination of scientific knowledge (Zhao et al., 2024). In contrast, technological teams focus on patents as their primary achievements, with technological innovation as their core objective (Ardito et al., 2021). In science-technology teams, academic inventors who hold dual roles as both paper authors and patent inventors bridge the gap between scientific discovery and technological application. This dual role is significant for analyzing the knowledge correlation and interaction mechanisms between science and technology (Xu et al., 2023). Therefore, it is necessary to break the limitations of less classifying analysis of scientific research teams, and provide more accurate cooperation strategy suggestions for different types of research teams from a differentiated perspective (Yoo et al., 2024).

According to the existing research gaps, this paper classifies research teams based on their unique roles within the innovation chain, and utilizes the CART decision tree algorithm to thoroughly explores the pathways for enhancing their innovation performance through network embeddedness. More specifically, we explored the following three questions: RQ1: What are the role that structural embeddedness, relational embeddedness and content embeddedness play in improving team innovation performance? RQ2: Are there significant differences in innovation performance improvement paths among scientific teams, technological teams, and science-technology teams? RQ3: What combination of embeddedness characteristics should different types of research teams focus on to enhance the innovation performance of research teams? This study contributes to the existing literature in several ways. First, we constructe a framework that integrates structural embeddedness, relational embeddedness and content embeddedness to explain team innovation performance. Second, by employing data-driven machine learning methods, this paper analyzes the complex nonlinear relationships and multi-factor combinations influencing the innovation performance of research teams. Third, this study proposes the multiple pathways to enhanced innovation performance in different types of research teams, which has important implications for different research teams in pharmaceutical to effectively manage network resources and facilitate high innovation, and are expected to be extended to other areas in future studies.

The article is structured as follows. The “Related Works” section introduces the research background of the study; “Data and Methods” section is then a description of the data and the method of analysis; the Results and Discussion section discusses the results; In the last section, the conclusion is presented, and policy suggestions are outlined.

Related Works

Network embeddedness

Network embeddedness theory characterizes the interactive relationship between social networks and individual organizations, emphasizing that individuals, groups, or organizations are embedded in social networks, where their behaviors and activities are influenced by others in the network (Schweitzer et al., 2022). With the rapid development of information technology, network embeddedness has become a crucial mechanism for linking individuals within and beyond teams, facilitating information exchange and knowledge sharing (Deng et al., 2023), and its influence on the innovation performance of teams has become increasingly evident (Ardito et al., 2021). In this context, this study adopts the perspective of network embeddedness to analyze the pathways for enhancing the innovation performance of three types of scientific research teams, focusing on the three dimensions of structural embeddedness, relational embeddedness, and content embeddedness.

Structural embeddedness and team innovation performance

Structural embeddedness is related to the degree of an actor's resource acquisition through their network position within the value network (Yan et al., 2019; Song et al., 2024; Deng et al., 2023), which includes central position, intermediary position, and other critical elements such as network density (Rowley et al., 2000). Central position is assessed using eigenvector centrality (Patel et al., 2024), which increases when connected to other key members who occupy significant positions within the network (Wasserman and Faust, 1994). Centrally located members in a network possess greater prominence and access to higher-quality information resources. They are more likely to receive insights from a wide range of individuals, which increases their potential for innovation (Burt, 2018; Sparrowe et al., 2001; Wasserman and Faust , 1994; Brass, 1984). The intermediary position is assessed through the concept of structural holes in a collaboration network, playing a crucial controlling role in connecting different subgroups (Nordt et al., 2024). Intermediary position in the network provide a supplementary resource exchange channel for the relatively dispersed network. A node located in a structural hole acts as the "bridge" and obtains a competitive advantage through information and control advantages (Tortoriello, 2015), which has a significant positive impact on team innovation performance (Pullen et al., 2012).

Relational embeddedness and team innovation performance

Relational embeddedness reflects the characteristics of the relationships between individuals in social network (Wang et al., 2024). Tie strength is one of the key indicators that reflects the embedding level of relationships (Deng et al., 2023; Wang, 2016). A commonly adopted approach to measuring the tie strength of cooperation is assessing the connections between nodes in a collaborative network (Schlattmann, 2016). Granovetter (1973) first divided collaborative ties into strong and weak ties. The "strong ties" refers to the multiple social relationships with high frequency, mutual trust and information sharing among network nodes while the

"weak ties" refers to a single social relationship in which the network nodes are less connected and less frequent (Granovetter, 1973). Liu et al. (2019) observed several nodes with high connection strength in the collaboration network, and further studied how do super-partnerships influence productivity and influence (Liu et al., 2024). Petersen (2015) similarly quantified the impact of nodes with strong partnerships and found that they contribute to produce above-average research output. Previous studies have shown that cooperative tie strength is closely related to the improvement of technological innovation performance (Lee, 2010). However, there is still debate about whether strong or weak relationships are more effective in promoting innovation performance. Coleman et al. argue that strong relationships are more conducive to innovation performance because they facilitate the acquisition of social capital, frequent communication, close contact, and mutual trust, all of which promote efficient information and knowledge sharing (Coleman et al., 1988). In contrast, Granovetter and Singh et al. suggest that weak relationships, with their heterogeneous knowledge and resources, are more conducive to generating new knowledge and effectively enhancing innovation performance (Granovetter, 1983; Singh, 2000).

Content embeddedness and team innovation performance

Network content embeddedness refers to the subjective innovation resources that assist the innovation subject in carrying out innovation activities and enhancing output. This characteristic primarily examines how the subjective characteristics of team members, as embedded within a network, influence the overall outcomes of the team (Zhou et al., 2024). The indicators used to measure content embeddedness include academic age diversity, research topic diversity, and knowledge interdependence. Academic age diversity refers to the variation in the academic ages of team members (Zou et al., 2023). This diversity may bring about different perspectives and problem-solving strategies, making the team more adaptable and innovative in response to evolving knowledge structure and societal needs (Sheng et al., 2024), thereby enhancing overall innovation performance (Dong et al., 2018). Research topic diversity reflects the degree of differentiation in the research interests of team members. Members with diverse research topics contribute a wealth of knowledge (Sheng et al., 2024), which enhances complementarity among team members, integrates various cognitive approaches, and accelerates scientific research and technological innovation (Chien et al., 2021). Empirical studies have demonstrated that the diversity of information and knowledge within a team significantly enhances its innovation performance (Schubert et al., 2020). Knowledge interdependence measures the degree of close connections and interactions among knowledge units in past knowledge combinations (Boxu et al., 2022). A shared language and routines, based on similar knowledge structures, reduce the time and effort needed for searching and communication, facilitating the sharing and recombination of diverse knowledge and lowering search and communication costs (Hansen and Haas, 2001). Moreover, individual members become more familiar with each other's knowledge bases, allowing them to reach a consensus on the direction of the search process. This familiarity facilitates the

communication, sharing, and transfer of tacit knowledge, easing interactions among members and promoting effective collaboration (Jin et al., 2015; Schmidt et al., 2022).

Data and methods

This study focuses on the non-linear correlation between network embeddedness and the innovation performance of various research teams. The basic process is shown in Fig. 1, which primarily encompasses three steps: (1) The Clique Percolation Method (CPM) is employed to identify scientific teams, technological teams and science-technology teams. (2) A model according to structural embeddedness - relational embeddedness - content embeddedness anchored in network embeddedness theory is built. Concurrently, team innovation performance is divided into novelty and impact to be measured. (3) Taking network embeddedness as conditional attributes, and network innovation performance as decision attributes, CART decision tree is employed to extract the decision rules affecting network innovation performance, and further reveal the performance improvement path of different research teams.

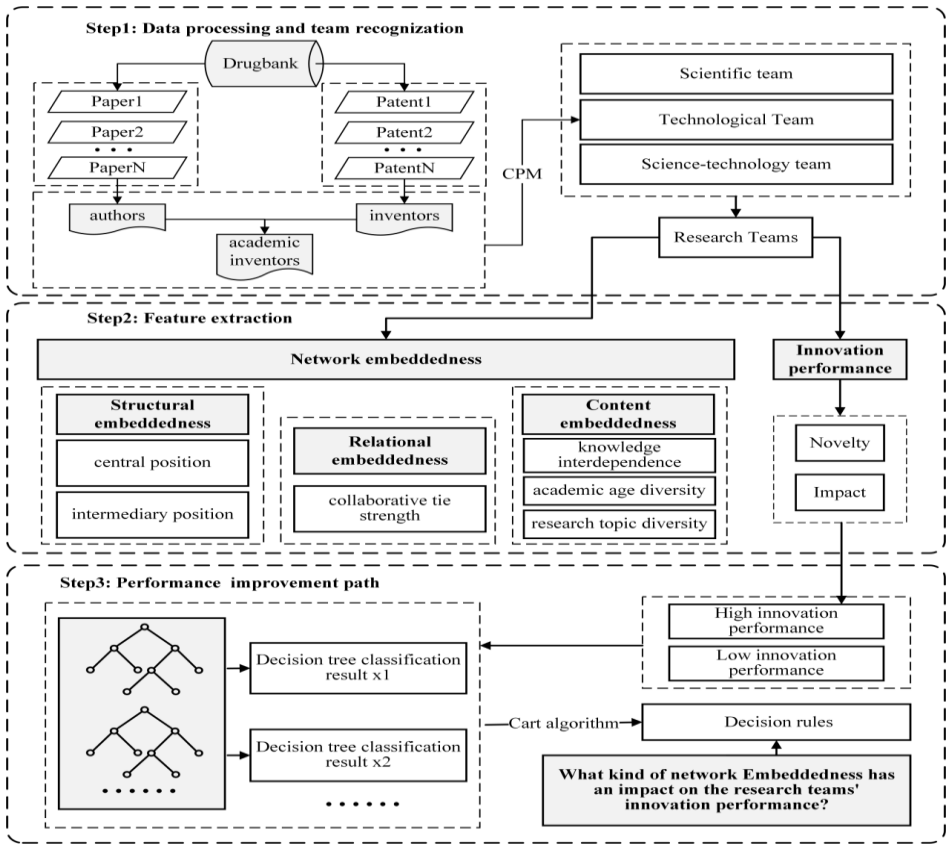


Figure 1. Research framework.

Data Sources

This study selects data from the field of pharmaceutical as an empirical research direction. The pharmaceutical field is grounded in science and driven by the interaction and collaboration among scholars with complementary knowledge, capabilities, and resources, which results in the production of relevant papers and patent information (Ashley, 2015). The dataset is derived from DrugBank (https://www.drugbank.ca/), a comprehensive, freely accessible online database in this field. Due to intensive science-based innovation embodied in drugs, related scientific publications and patents are explicitly linked to the resulting drugs (Strattonet al., 2024; Cantner & Rake, 2014).

This study builds a comprehensive repository by integrating multiple data sources related to drugs, scientific papers, and patents. On 1st November 2019, the DrugBank dataset is first obtained in XML format and imported into a MySQL database, generating a collection of 13,339 drug entries, 10,355 scientific papers, and 5,932 patents (Xu et al., 2021). Detailed information for the associated scientific papers (including authors, titles, abstracts, publication years, and DOIs) is obtained from the PubMed database using each paper's PubMed ID (PMID). In addition, citation counts for these papers are retrieved from the Scopus database using each paper's DOI, and references to the papers are collected from the Web of Science database. For patents, information is retrieved from the European Patent Office (EPO) database using patent publication numbers, yielding details such as inventors, titles, abstracts, publication dates, and publication years; patent application dates and patent citation counts are also obtained from the Incopat patent database. Ultimately, a total of 191,744 references for the papers and 73,030 citations for the patents are compiled as data sources for this study.

Team Recognition

Utilizing the DrugBank dataset, which links drug-related scientific publications and patents, this study identifies scientific teams, technological teams, and science-technology teams. The identification process follows a three-step approach: constructing collaboration networks, applying the faction filtering method to select teams, and matching teams across successive time periods.

1) Construction of Collaboration Networks. To ensure the reliability of team identification, this study excludes cases where no publications are produced for more than three consecutive years, thereby obtains the time span for authors range from 1943 to 2019, inventors from 1986 to 2019, and academic inventors are 1983-2019. Collaboration within teams typically occurs over a limited period, concentrated within three to five years (Guan et al., 2016). Accordingly, this study adopts a five-year sliding time window, with annual rolling updates, to construct separate collaboration networks for authors, inventors, and academic inventors. Each network represents the dynamic collaborative relationships within its respective group.

2) Faction Filtering Method for Team Identification. The Clique Percolation Method (CPM) is employed to identify scientific teams, technological teams, and science-technology teams within the author collaboration network, inventor

collaboration network, and academic inventor collaboration network, respectively. The process is implemented as follows: First, to enhance the consistency of identifying team members across time-slice collaboration networks, the k-value is set to 3, following prior studies. Second, nodes with fewer than two collaborations with other researchers are classified as inactive nodes. To filter out members with positive collaborative relationships, the edge weight threshold is set to 2. Lastly, the Python-based NetworkX toolkit is utilized to calculate the attributes of collaboration networks and implement the CPM algorithm for team identification. This approach generates the annual distribution of various research team types, as shown in Figure 2.

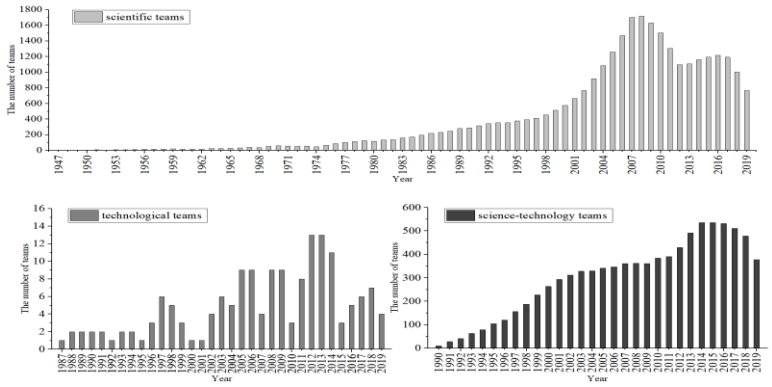


Figure 2. Distribution of number of scientific and technological teams distributed in different years.

3) Team Matching Across Adjacent Time Periods. To maintain continuity among the three distinct types of teams within the time series, this research employs community overlap metrics to monitor and align teams across successive time intervals (Liu et al., 2019). Specifically, joint networks are established from collaboration networks within consecutive temporal windows, and the team overlap coefficient *C* is employed to correlate teams at time *t* and *t+1*, thereby facilitating the identification of the same team across various periods. The calculation formula is shown in Equation (3-1). For a given team, the average overlap coefficient α is used to measure the stability of the team at time *t*. The calculation formula is presented in Equation (3-2).

$$C(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3-1}$$

$$\alpha = \frac{\sum_{t=t_0}^{t_{max}-1} C(t,t+1)}{t_{max} - t_0 - 1} \tag{3-2}$$

Here, $A \cap B$ refers to the number of identical members in team A and team B at adjacent moments, and $A \cup B$ refers to the total number of team members in team A and team B at adjacent moments. t_0 indicates the birth of the team, while t_{max} represents the last step before the team disappears.

A threshold value of 0.35 is established for the overlap coefficient to assess team continuity. Statistical analysis of the average overlap coefficient for each team type at time t reveals that the mean overlap coefficients for scientific teams, technological

teams, and science-technology teams all exceed 0.55. This finding indicates that team members experience relatively little change across different periods, and have a certain degree of stability and continuity. Eventually, 6,037 evolutionary sequences are identified for scientific teams, 1,442 for technological teams, and 66 for science-technology teams. To construct a static sample for further analysis, the members of each research team's evolutionary sequence are merged across all time slices, with duplicate members removed. This process results in a final dataset comprising 6,031 scientific teams, 1,378 technological teams, and 66 science-technology teams.

Variable definition and measurement

Dependent variables

Building upon the research conducted by Lee et al. (Lee et al., 2015), this paper jointly measures the innovation performance of research teams in terms of novelty and impact. Novelty refers to the ability of an output to provide a distinctive perspective, methodology, or technological advancement (Uzzi et al., 2021), while impact denotes the actual value of an innovative result in academia or practice (Cepoiet et al., 2023). Regarding the measurement of novelty, various strategic approaches have been proposed due to the differences in assessment between papers and patents.. For instance, Uzzi et.al (2021) argue that scientific papers citing an unusual combination of journals in their references can be considered to represent relatively more novel knowledge. In patents, most scholars have used backward citations to measure the novelty of a technology (Zhao et al., 2019). Therefore, this paper measures the novelty of papers and patents separately, and use forward citations to calculate their impact (Breitzman, 2021).

(1) Novelty

Based on the measure of novelty proposed by Lee et al. (2015), this paper calculates the novelty index score for each individual paper within a team, and subsequently computes the novelty score at the team level. Here, paper novelty refers to the rarity or innovation of the pairwise combinations of cited references that have been previously cited in the paper. The specific steps are as follows:

1) For each paper, all paired reference combinations are identified, and the corresponding pairs of journals associated with each reference combination pair are recorded. Second, to ensure data robustness, a three-year time window from *t-2* to *t* is used. The pairs of journal combinations corresponding to the papers published within this time window are summarized to generate the set U_t. And the commonness value is calculated according to Equation 3-3.

$$Commonness_{ijt} = \frac{observed\ number\ of\ pairs_{ijt}}{expected\ number\ of\ pairs_{ijt}} = \frac{N_{ijt}}{\frac{N_{it}}{N_t} \times \frac{N_{jt}}{N_t} \times N_t} = \frac{N_{ijt} \times N_t}{N_{it} \times N_{jt}} \quad (3\text{-}3)$$

Here, N_{ijt} represents the number of times the journal pair (i, j) appears in U_t, N_{it} represents the number of journal pairs in U_t that include journal i, N_{jt} represents the number of journal pairs in U_t that include journal j, and N_t represents the total number of journal pairs in U_t.

2) For a paper published in year t, the commonness values of all its journal pairs are ranked in descending order. To reduce noise and improve the reliability of this measure, the 10th percentile is then taken as the paper's commonness value. This value is transformed using the natural logarithm to approximate a normally distributed variable, resulting in a non-negative integer that represents the paper's novelty index.

3) Finally, the novelty innovation performance of each team is assessed based on the novelty index of papers produced by the team. The formula is as follows:

$$C(team) = \frac{\sum C(i)}{N} \quad (3\text{-}4)$$

Here, $C(team)$ represents the team novelty, $C(i)$ represents the paper novelty, and N represents the team size.

According to research conducted by Zhao et al. (2019), a lower number of backward citations for a patent indicates that it relies less on prior art and possesses a higher degree of technical novelty. Additionally, some scholars propose that a reduction in backward citations signifies enhanced uniqueness and significant differences compared to existing patents. Consequently, this study selects the number of backward citations of patents as a proxy variable for technological novelty. This variable takes non-negative integer values and is classified as a count variable.

(2) Impact

The impact of a team is assessed using the forward citation method (Lee et al., 2015). This paper defines a high-impact paper as one that ranks in the top 1% (i.e., the 99th percentile) of its citation distribution. The specific identification steps are as follows: 1) Rank all papers by their citation counts in descending order. 2) To avoid potential misidentification of highly cited papers due to the use of a short citation time window (Zhang et al., 2021), a five-year moving window is applied. Papers ranked in the top 1% from year t-5 to year t are classified as high-impact papers. 3) Represent the impact status of each paper using a binary variable: assign a value of 1 if the paper is considered highly impactful, and 0 otherwise. 4) Count the number of high-impact papers for each team, then divide the total by the team size to calculate the team's impact score.

Independent variables

This study draws on previous research by scholars regarding the characteristics of collaborative networks and their impact on innovation performance. It adopts a network embeddedness perspective to identify six representative indicators across three dimensions: structural embeddedness, relational embeddedness, and content embeddedness, which collectively characterize the features of research team collaboration networks. Structural embeddedness pertains to central and intermediary position. Relational embeddedness includes the strength of collaborative ties. Content embeddedness is further divided into three aspects:

academic age diversity, research topic diversity, and knowledge interdependence. The specific measurement methods for these indicators are as follows.

(1) Central position

The central position of a team within a network is primarily assessed using eigenvector centrality (Patel et al., 2024). Eigenvector centrality reflects the extent of a participant's connections within the network and serves as a crucial important indicator for evaluating the impact of team members. Specifically, a member's eigenvector centrality increases when they are connected to other key members who occupy significant positions within the network. The formula is as follows:

$$EC_i = c \sum_{j=1}^{n} a_{ij} \cdot x_j \quad (3\text{-}5)$$

Here, *c* represents *a* constant, a_{ij} denotes the adjacency matrix of the network, and j represents the adjacent nodes of node *i*. Therefore, for a given node *i*, its eigenvector central position is proportional to the sum of the central position values of all nodes connected to it. Finally, the central position of the team is determined by dividing the total central position of all members by the size of the team.

(2) Intermediary position

The intermediary position is typically assessed using metrics related to structural holes within the collaborative networks of research teams across organizations. Burt's structural hole metrics include effective size, efficiency, and the constraint coefficient. This study utilizes the constraint coefficient to quantify intermediary position, with the formula as follows:

$$C_{ij} = \left(p_{ij} + \sum\nolimits_q p_{iq} p_{jq} \right)^2 \quad (3\text{-}6)$$

$$C_i = \sum\nolimits_i C_{ij} \quad (3\text{-}7)$$

Here, j represents all nodes connected to the ego node, and q represents the third node other than i or j. $p_{iq}p_{jq}$ denotes the strength of the relationship between node i and node j, and the product term represents the redundancy between the ego node and node j. p_{iq} indicates the proportion of the ego node's resources invested in its relationship with node i, and p_{jq} represents the strength of the relationship between node i and node j. Finally, the intermediary position of the team is calculated by dividing the total number of member intermediary positions by the team size.

(3) Collaborative tie strength

Collaborative tie strength refers to the degree of connection strength among all nodes within a network. The collaborative tie strength of a team is determined by constructing the teamwork network and measuring the degree of connectivity among all nodes within the network. The calculation formula is as follows:

$$S(G) = \frac{\sum v_{ij}}{E} \quad (3\text{-}8)$$

Here, $\sum v_{ij}$ denotes the sum of collaboration frequencies among the network nodes, and E represents the total number of edges in the network.

(4) Academic age diversity

Academic age was quantified by assessing the duration of each team member's academic career, denoted as Y_l for the year of publication of the researcher's most recent paper or patent, and Y_0 for the year of publication of the researcher's initial paper or patent. The metrics for academic age diversity were evaluated utilizing Simpson's index (Mao et al., 2024), which was computed according to the following formula:

$$P_i = Y_l - Y_0 + 1 \quad (3\text{-}9)$$

$$H = 1 - \sum_{i=1}^{n} P_i^2 \quad (3\text{-}10)$$

where n denotes the total number of categories and P_i denotes the percentage of members of the team in category i The higher the value of H, the greater the diversity.

(5) Research topic diversity

First, the abstracts and author order of the team's publications (papers or patents) are standardized. Next, the ATcredit model (Xu et al., 2021) is used for topic modeling, and the cosine similarity of topics is calculated based on the two output files, "researcher-topic" and "topic-word probabilities" from the model. This study adopts the diversity indicator calculation and application methods outlined by Leydesdorff et al. (Leydesdorff, 2018). Team research topic diversity is represented by $RTD_Team_{m,i}$, indicating the research topic diversity of member m within the *i-th* research team. The specific calculation method is shown in Equation 3-11.

$$RTD_Team_{m,i} = \left(\frac{n_{m,i}}{N_i}\right) * \left(1 - Gini_{m,i}\right) * \sum_{\substack{i,j \\ i \neq j}}^{n_{m,i}} \frac{d_{ij}}{n_{m,i}(n_{m,i} - 1)} \quad (3\text{-}11)$$

$$Gini_{m,i} = \frac{\sum_{i=1}^{n_{m,i}} (2i - n_{m,i} - 1)x_i}{n_{m,i} \sum_{i=1}^{n} x_i} \quad (3\text{-}12)$$

Here, N_i represents the total number of available topics for the publications of the *i-th* team; $n_{m,i}$ denotes the number of topics assigned to member m in the *i-th* team; $1 - Gini_{m,i}$ represents the evenness of topic diversity in the publications of member m in the *i-th* team; d_{ij} indicates the difference between topics i and j, calculated using "1 - cosine similarity." The coefficient $Gini$ is used to measure the imbalance in frequency distribution values, with its calculation shown in Equation 3-12, where x_i represents the *i-th* observation. Finally, the research topic diversity of the team is determined by dividing the sum of its members' research topic diversity by the team size.

In particular, considering that the innovations of science-technology teams cover both papers and patents, the diversity of research topics of such teams is further calculated after weighting papers and patents using the entropy weighting method (Bai et al., 2020).

(6) Knowledge interdependence

Knowledge interdependence measures the degree of close connection and interaction between knowledge units in past knowledge combinations. This paper adopts the

method and calculation principles proposed by Fleming and Sorenson (Fleming et al., 2001) to measure the knowledge interdependence of three different types of teams.

In the context of scientific teams, the medical subject terms derived from the DrugBank dataset effectively represent the biomedical topics addressed in the respective papers. Therefore, the knowledge interdependence of scientific teams is measured using the following two steps. First, for a paper p containing N biomedical subject terms k, the ease of combination of the biomedical subject terms P_E_k is calculated, with the specific calculation method shown in Equation (3-13). Second, the ease of combining medical subject terms P_E_k for all N biomedical subject terms in the paper p is summed, and the reciprocal of the arithmetic mean of the summed value is taken to obtain the knowledge interdependence of the paper p. The specific calculation method is shown in Equation (3-14). Finally, the knowledge interdependence of the team is determined by dividing the total knowledge dependence of the scientific team members' published papers by the total number of papers.

$$P_E_k = \frac{\begin{array}{c}\text{In papers published earlier than } p,\\ \text{the number of medical subject words that}\\ \text{have a combinatorial relationship with medical subject words } k\end{array}}{\begin{array}{c}\text{The number of papers published earlier than } p \text{ that}\\ \text{contain the medical subject word } k\end{array}} \quad (3\text{-}13)$$

$$P_IND_p = \frac{N}{\sum_{k=1}^{N} P_E_k} \quad (3\text{-}14)$$

For technological teams, the patent IPC classification codes derived from the DrugBank dataset effectively represent the relevant technical domains and subjects associated with the patents. Consequently, two subsequent steps will be implemented to assess the knowledge interdependence among technology teams. First, for a patent q containing M IPC classification codes, the ease of combination of the IPC classification codes $\mathrm{Q_E_k}$ is calculated, with the specific calculation method shown in Equation (3-15). Second, the ease of combination for all M IPC classification codes Q_E_k in the patent q is summed, and the reciprocal of the arithmetic mean of the summed value is taken to obtain the knowledge interdependence of the patent q. The specific calculation method is shown in Equation (3-16). Finally, the knowledge interdependence of the team is obtained by dividing the total knowledge dependencies of the technological team members' patent applications by the total number of patents.

$$\mathrm{Q_E_k} = \frac{\begin{array}{c}\text{In patents filed earlier than } q,\\ \text{The number of IPC class numbers that}\\ \text{have a combination relationship with the IPC class number } k\end{array}}{\begin{array}{c}\text{The number of patents filed earlier than } q \text{ that}\\ \text{contain IPC class number } k\end{array}} \quad (3\text{-}15)$$

$$Q_IND_q = \frac{M}{\sum_{k=1}^{N} Q_E_k} \quad (3\text{-}16)$$

For science-technology teams, the process of calculating knowledge interdependence is similar to measuring of research topic diversity. Knowledge interdependence must be quantified based on the respective publications and patents produced by these teams. Ultimately, the knowledge interdependence of science-technology teams is calculated by applying the entropy weighting method to the aforementioned results.

Research method

To explore the pathways for improving the innovation performance of research teams under various feature combinations, this study employs the CART (Classification and Regression Trees) decision tree algorithm as the primary analytical tool. The CART algorithm is a binary decision tree-based method for classification and regression. Its results provide high interpretability and robustness, and it has been proven to be an effective model for uncovering the impact pathways of team innovation performance (Zhang et al., 2023).

The rationale for choosing the CART algorithm is as follows. First, the CART algorithm recursively performs binary splits, enabling it to effectively identify significant features and conditions within the dataset, thereby generating clear and precise decision rules. These rules are represented as node combinations along each path of the decision tree, visually presenting the pathways for improving innovation performance across different types of research teams (Lyu et al., 2021). Second, during the attribute splitting process, the CART algorithm prioritizes the attribute that results in the lowest Gini coefficient, ensuring high sample purity and enhancing both the accuracy and interpretability of the model. The formula for the Gini coefficient is shown in Equation (3-17).

$$Gini(P) = \sum_{i=1}^{N} P_i(1 - P_i) \quad (3\text{-}17)$$

Finally, in comparison to other decision tree algorithms such as ID3, C4.5, and C5.0, the CART algorithm generates more concise decision trees. It minimizes the generation of excessive branches (Lyu et al., 2021), which facilitates the interpretation and application of decision rules.

To validate the effectiveness of the CART algorithm, this study divides the data into training and test samples in an 8:2 ratio. In addition to the CART model, various baseline models are developed, including the C4.5 decision tree model, random forest classifier, and gradient boosting classifier. The performance of these models on the test set is compared to select the optimal model, further confirming the advantages of the CART algorithm.

Moreover, the study employs grid search (GridSearchCV) to thoroughly explore and evaluate the hyperparameter combinations of the models, aiming to identify the optimal configuration for improving model performance. After optimizing the models through grid search, the accuracy of all models exceeds 0.6, indicating that

more than 60% of the samples are accurately predicted. In the scientific team model, the CART decision tree achieves accuracy rates of 0.87 for novelty and 0.92 for impact, respectively. In the technology team model, the CART decision tree also performs exceptionally well, with accuracy rates of 0.82 for novelty and 0.89 for impact. In the scientific-technological team model, accuracy rates reach impressive levels of 0.85 for novelty and 0.95 for impact.

Results and discussion

Analysis of Network Embeddedness Characteristics of Research Teams

Based on the aforementioned variable definitions and model framework, combined with the identified team samples, this study measures the network embeddedness characteristics of three research teams. Basic information of the variables is shown in Table 1. Among them, the network embeddedness characteristics represent the average value of various types of research team network embeddedness characteristics. The average value of each variable is shown in Table 1. From the perspective of research team network embeddedness characteristics, there are certain differences across team types, further validating the scientific and rationality of team definition and identification, as well as the significance of exploring the pathways for improving innovation performance in different types of teams.

The average value of the intermediary position in science-technology teams is the highest compared to the other two types of teams, indicating that members of science-technology teams more frequently play the role of brokers in the collaboration network. Compared to other types of research teams, members of this team have a dual identity and link the academic and technological fields to a greater extent, which explains their higher intermediary position. The members of technological teams have greater research topic diversity, which may stem from the diversified knowledge of the technological team members and the complex innovation goals within the team (Liu et al., 2024).

Table 1. Research team collaboration network embedded information.

Team category	Number of teams	Central position	Intermediary position	Collaborative tie strength	Academic age diversity	Knowledge interdependence	Topic diversity
Scientific teams	6,031	0.01	0.12	1.07	0.20	0.16	0.75
Technological teams	1,378	0.36	0.09	2.21	0.30	0.39	0.83
Science-technology teams	66	0.01	0.55	3.35	0.62	0.23	0.18

Correlation Analysis

To highlight the need for machine learning in analytical work, Figure 3 shows the correlations between network embeddedness characteristics and innovation performance across three types of research teams. Overall, the correlation coefficients between central position, intermediary position, collaborative tie strength, academic age diversity, research topic diversity, knowledge interdependence, and research team innovation performance are generally below 0.6, with most statistical test results being insignificant. This indicates that research team innovation performance is not driven by any single factor, but rather results from a complex interplay of multiple dimensions. To explore these nonlinear and multifaceted relationships, this study employs the CART decision tree algorithm to uncover effective pathways to high innovation performance for different team types.

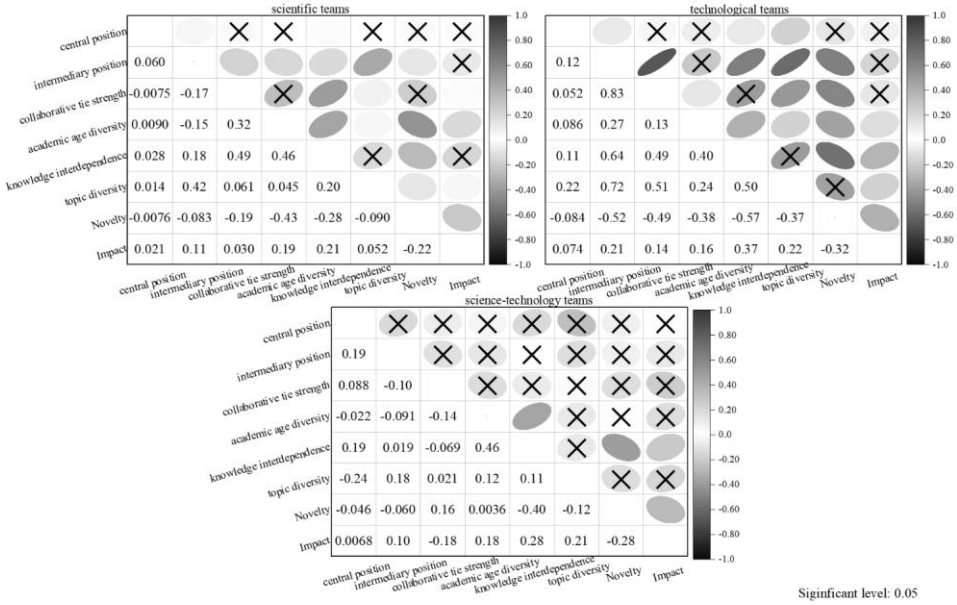


Figure 3. The correlation between network embeddedness characteristics and innovation performance of three types of research teams. (A cross indicates that the correlation is not significant; Different levels of gray circles indicate the strength of correlation, the darker the color, the stronger the correlation; The direction of the gray circle indicates the positive and negative correlation, with the right being positive and the left being negative).

Decision Rules for Research Team Innovation Performance

To further explore the impact of collaborative network embeddedness characteristics on innovation performance, six characteristics—central position, intermediary position, collaborative tie strength, academic age diversity, research topic diversity, and knowledge interdependence—are used as the six condition attributes for decision rules. Innovation performance is assessed through two outcome attributes: novelty and impact. The CART decision tree algorithm is applied to extract rules for each of the six network embeddedness characteristics, thereby revealing their relationship with innovation performance. The resulting decision rules are shown in Tables 3 and 4. Among them, support refers to the proportion of samples that meet a given rule, reflecting its coverage. A higher support indicates that the current decision rule can explain more cases. Confidence refers to the proportion of correctly predicted positive outcomes among those that satisfy the rule conditions—essentially, the rule's predictive reliability. Higher confidence implies greater reliability of the rule in predicting outcomes.

Table 3. Decision rules table-novelty.

Team category	Structural embeddedness		Relational embeddedness	Content embeddedness			Performance level	Support	Confidence
	Central position	Intermediary position	Collaborative tie strength	Academic age diversity	Knowledge interdepend-ence	Topic diversity			
Scientific teams		<=0.16		<=0.07			H	40.00%	92.00%
		>0.16		<=0.07			L	24.00%	79.00%
				>0.07	<=0.02		H	16.00%	69.00%
				>0.07	>0.02		L	24.00%	69.00%
Technological teams					<=0.04		H	47.00%	98.00%
					(0.04,0.06]		H	36.00%	74.00%
	<=0.01				>0.06		H	31.00%	60.00%
	>0.01				>0.06		L	15.00%	83.00%
Science-technology teams	<=0.03						H	12.00%	100.00%
	>0.03		<=0.68	<=0.71			L	36.00%	64.00%
	>0.03		<=0.68	>0.71			H	23.00%	73.00%
	>0.03		>0.68				H	8.00%	100.00%

*H represents high performance level and L represents low performance level.

Table 4 .Decision rules table-impact.

Team category	Structural embeddedness		Relational embeddedness	Content embeddedness			Performance level	Support	Confidence
	Central position	Intermediary position	Collaborative tie strength	Academic age diversity	Knowledge interdepend-ence	Topic diversity			
Scientific teams					<=0.03		L	30.00%	64.00%
					(0.03,0.09]		L	45.00%	89.00%
	<=0.01				>0.09		L	14.00%	78.00%
	>0.01				>0.09		H	15.00%	88.00%
Technological teams		<=0.07		<=0.32			L	25.00%	98.00%
		>0.07		<=0.32			L	44.00%	90.00%
				>0.32	<=0.39		L	13.00%	91.00%
				>0.32	>=0.39		H	10.00%	71.00%
Science-technology teams			<=0.19	<=0.65			L	12.00%	88.00%
			<=0.19	>0.65			H	15.00%	60.00%
			>0.19		<=0.33		L	39.00%	96.00%
			>0.19		>0.33		L	12.00%	62.00%

*H represents high performance level and L represents low performance level.

A comprehensive analysis yields several key insights: (1) The confidence levels of the decision rules range from 60% to 100%, indicating good model fit and strong interpretability. (2) There are significant differences in the impact of central position, intermediary position, collaborative tie strength, academic age diversity, research topic diversity, and knowledge interdependence on innovation performance across different types of research teams. This variation in decision rules highlights the necessity for classifying and modeling different team types separately. (3) A comparison of the two dimensions of innovation performance shows that a greater proportion of teams achieve high performance in the “novelty” dimension than in “impact”, supporting the need to analyze these dimensions independently.

Innovation Performance Improvement Path in Scientific Teams

Figures 4 and 5 present the decision trees for scientific teams, corresponding to the two dimensions of innovation performance: novelty and impact. Figure 4 identifies four decision rules associated with scientific team novelty. Among them, the rules “low academic age diversity → low intermediary position” and “high academic age diversity → low knowledge interdependence” emerge as two pathways for enhancing the novelty. Notably, academic age diversity serves as the root node, exhibiting the strongest relationship with scientific team novelty.

In scientific teams, low academic age diversity (≤0.07) suggests that members share similar academic trajectories. When the intermediary position is also lower (≤0.16), a relatively equal communication platform can be formed. Team members with similar academic ages can collaborate more effectively, sharing resources and

information rather than competing—contributing to a significant improvement in team's novelty innovation performance. In teams with higher academic age diversity (>0.07)—such as teams composed of senior, mid-career, and early-career researchers—a cross-generational and diversified knowledge exchange and innovation platform can be established. Within this context, low knowledge interdependence (≤0.02) helps break the "lock-in effect" and past knowledge combinations (Rishika et al., 2019), fostering openness to external knowledge sources within the collaborative network. Conversely, teams above this threshold are at risk of declining innovative performance.

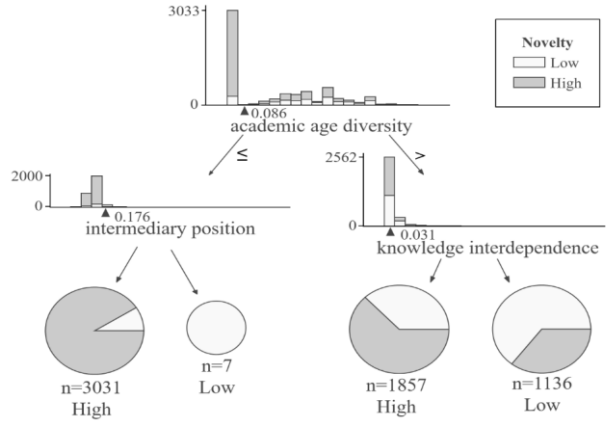


Figure 4. Decision tree of scientific teams-novelty.

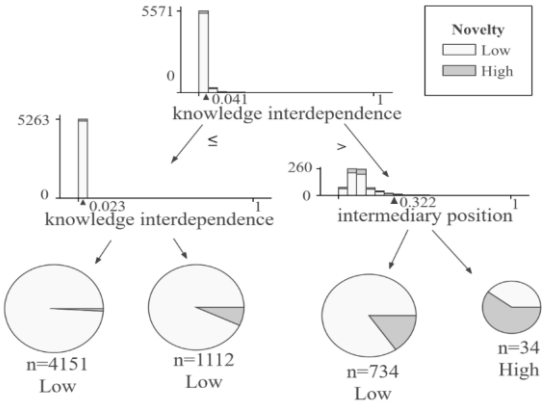


Figure 5. Decision tree of scientific teams-impact.

As shown in Figure 5, knowledge interdependence emerges as the root node, showing a strong relationship with innovation performance in terms of impact. In the case of low knowledge interdependence (≤0.09), the team's impact remains at a low level. The only decision rule supporting high impact in scientific teams is "high knowledge interdependence → high central position". In this scenario, members with high knowledge interdependence (>0.09) are embedded in a rich accumulation of past knowledge. Simultaneously, occupying a central position (>0.01) allows them to connect with authoritative nodes in the field. This dual advantage enables such

teams to generate novel insights by building on established expertise, thereby attracting greater academic visibility and citations—ultimately leading to significantly enhanced impact performance.

Innovation Performance Improvement Path in Technological Teams

Figures 6 and 7 illustrate the decision trees for technological teams, corresponding to the two dimensions of innovation performance: novelty and impact. In Figure 5, the decision tree for novelty in technological teams presents four decision rule, and knowledge interdependence appears as the root node. Two distinct pathways are identified for enhancing novelty performance: "low knowledge interdependence" and "high knowledge interdependence → low central position".

A lower degree of knowledge interdependence (≤0.06) indicates that members of technological teams face fewer constraints from established knowledge systems, enabling them to achieve higher novelty of outcomes. Under high knowledge interdependence (> 0.06), team members with low centrality (≤0.01) are less connected to core nodes. Independent research and development efforts by these peripheral or "marginal" scholars can reduce knowledge redundancy caused by over reliance on past expertise, further promoting innovation novelty within the team.

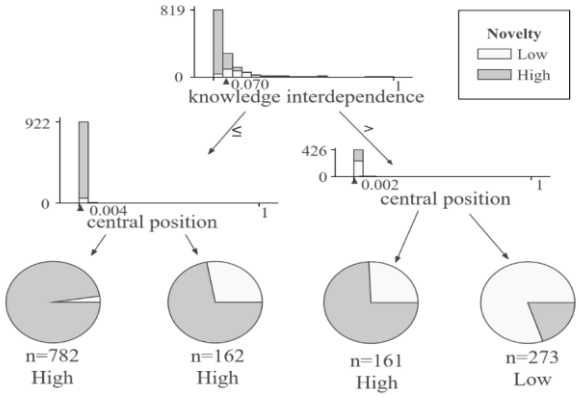


Figure 7. Decision tree of technology teams-impact.

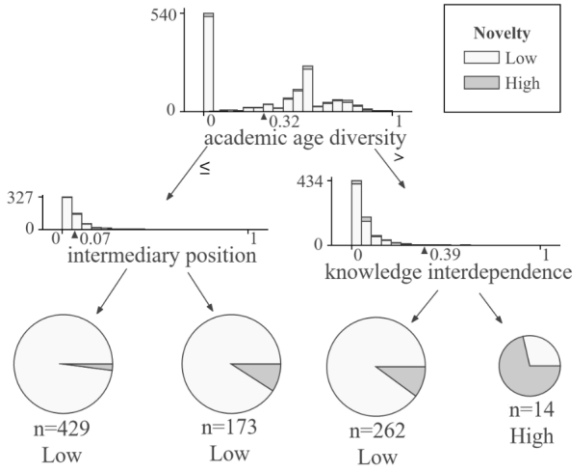


Figure 6. Decision tree of technology teams-novelty.

As illustrated in Figure 7, the decision tree for the impact dimension of technological teams includes four decision rules, with academic age diversity serving as the root node. When academic age diversity is low (≤ 0.32), teams generally exhibit consistently low levels of impact. Among the identified rules, only the rule "high academic age diversity → high knowledge interdependence" is associated with improved impact performance in technological teams. High academic age diversity (> 0.32) equips technological teams with a broader range of perspectives and approaches. Simultaneously, high knowledge interdependence (> 0.39) reflects a collective recognition for the team's technical knowledge base, enabling members to rapidly integrate cross-generational expertise—thereby significantly enhancing the team's overall impact.

Innovation Performance Improvement Path in Science-technology Teams

Figures 8 and 9 present the decision trees for science-technology teams, corresponding to the two dimensions of innovation performance: novelty and impact. As shown in Figure 7, the decision tree for the novelty dimension identifies four decision rules, three of which are associated with higher levels of innovation performance: "low central position" , "high central position → high tie strenth" and "high central position → low tie strenth → high academic age diversity".

In science-technology teams, academic inventors often operate within two distinct networks: one oriented toward academic research and the other toward patent activity. Centrality in one network is typically achieved at the expense of centrality in the other. A low central position (≤ 0.03) allows academic inventors greater autonomy in making strategic trade-offs, helping them delve into scientific frontiers and produce innovative research outcomes. When tie strength with high-centrality nodes (> 0.03) is high (> 0.68), academic inventors benefit from former shared language, practices, and frameworks, which allows for continued access to accumulated expertise and frontier insights, contributing to higher levels of novelty in research output. In contrast, when team members occupy high-centrality positions (> 0.3) but maintain weaker ties (≤ 0.68), their ability to mobilize and exchange resources with key nodes is limited. In such cases, extensive academic experience (> 0.71) enables the team to integrate diverse knowledge and technical skills, compensating for weaker connections. This facilitates an alternative route to innovation, overcoming constraints in resource exchange and supporting high-impact performance.

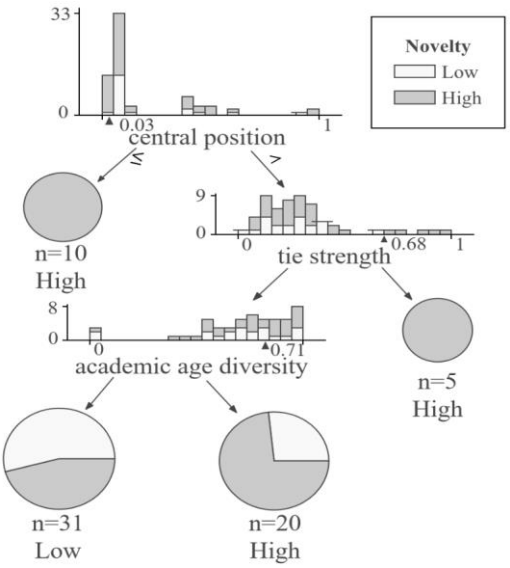


Figure 8. Decision tree of science-technology teams-novelty.

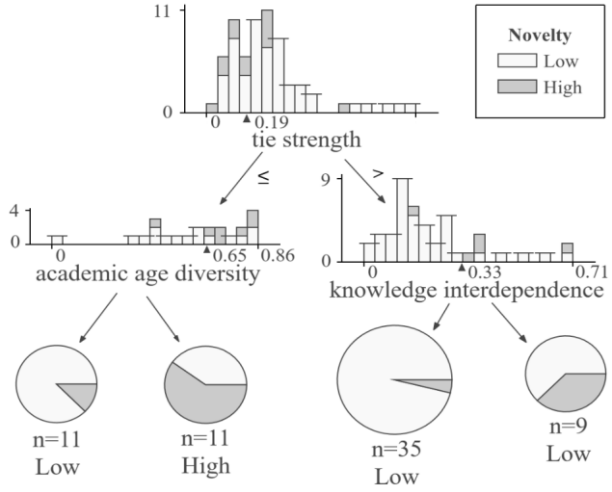


Figure 9. Decision tree of science-technology teams-impact.

The decision tree in Figure 8 shows four decision rules, with collaborative tie strength being identified as the root node, indicating its relevance to innovation performance related to impact for science-technology teams. Teams with high collaborative tie strength (>0.19) typically exhibit consistently low levels of impact. Only the decision rule "low tie strength → high academic age diversity" exhibits a significant correlation with higher levels of impact in science-technology teams.

Research indicates that high-intensity collaboration among academic inventors can be associated with increased knowledge sharing. However, it may also correlate with knowledge redundancy and potentially trigger collaboration conflicts that can challenge team cohesion and may be linked to reduced overall team impact. In contrast, academic inventors with lower collaborative tie strength (≤0.19) and higher academic age diversity (>0.65) are typically better able to avoid collaboration

conflicts. By focusing more on their own research fields, they can attract collaborators with varying academic backgrounds, enriching the openness of the network's knowledge and the diversity of its resources. This ongoing process of resource accumulation and enhancement can substantially elevate the impact of the team.

The final importance of the explanatory variables obtained from the model is presented in Table 5 and Table 6, respectively. Among the factors of "novelty" and "impact", which the model predicted to influence the innovation performance of the three types of research teams, knowledge dependence emerged as the most significant factor, with importance values of 1.000 and 0.751, respectively. This finding highlights that, among the various factors contributing to research teams' innovation performance, knowledge dependence exhibits the strongest correlation. It underscores the critical role of knowledge flow and the interdependence among team members in enhancing a team's innovative capacity. A high degree of knowledge dependence within a team enables members to leverage each other's expertise more effectively, fostering collaboration and driving the team's overall innovation performance.

Table 5. Characteristic importance of explanatory variables-novelty.

Team's category	Central position	Intermediary position	Collaborative tie strength	Academic age diversity	Knowledge interdependence	Topic diversity
Scientific teams	0	0.015	0	0.794	0.191	0
Technological teams	0.317	0	0	0	0.683	0
Science-technology teams	0	0	0	0	1.000	0

Table 6. Characteristic importance of explanatory variables-impact.

Team's category	Central position	Intermediary position	Collaborative tie strength	Academic age diversity	Knowledge interdependence	Topic diversity
Scientific teams	0	0.249	0	0	0.751	0
Technological teams	0.329	0	0	0	0.635	0.036
Science-technology teams	0	0	0.010	0	0.990	0

Conclusion

Key findings

Based on literature and patent data from the pharmaceutical field, this study identifies scientific teams, technological teams, and science-technology teams. Utilizing embeddedness theory, a framework is developed for analyzing innovation performance in research team collaboration networks, encompassing three dimensions: structural embeddedness, relational embeddedness, and content embeddedness. The CART decision tree algorithm is applied to investigate the synergistic pathways by which the characteristics of collaboration networks contribute to enhanced innovation performance across various types of research teams. The study yields several significant conclusions:

(1) The improvement of innovation performance within research teams is not a straightforward correlation of a single variable; rather, it is a complex process influenced by multiple dimensions and factors. Table 7 presents an overview of the results regarding the pathways to improved innovation performance for various types of teams. Overall, within different categories of research teams, the combined effects of central position, intermediary position, collaborative tie strength, academic age diversity, and knowledge interdependence on innovation performance show significant differences. The decision rules vary considerably, underscoring the scientific necessity of classifying and modeling different types of research teams for further discussion. Knowledge interdependence is a key indicator influencing the innovation performance of different types of research teams. Referring to Table 7, low knowledge interdependence is identified as a key factor in enhancing novelty, whereas high knowledge interdependence is essential for improving impact.

Table 7. Summary of the decision rules.

Team's category	Novelty	Impact
Scientific teams	"academic age diversity↓→ intermediary position↓", "academic age diversity↑→knowledge interdependence↓"	"knowledge interdependence↑→ central position↑"
Technological teams	"knowledge interdependence↓", "knowledge interdependence↑→ central position↓"	"academic age diversity↑→ knowledge interdependence↑"
Science-technology teams	"central position↓", "central position↑→ tie strength↑", "central position↑→tie strength↓→ academic age diversity↑"	"tie strength↓→ academic age diversity↑"

(2) In scientific teams, when occupying low intermediary positions and academic age diversity, team members are more likely to achieve a significant increase in the novelty innovation performance. Meanwhile, teams with high academic age diversity and low knowledge interdependence can experience high innovation performance. This independence, coupled with intergenerational collaboration, often leads to enhanced novelty innovation performance. Regarding impact, teams with high knowledge interdependence and central position benefit from both accumulated expertise and strong connections to influential actors within their network. These dual advantages significantly enhance the overall innovation performance.
(3) In technological teams, those with lower knowledge dependence are less dependent on past knowledge combinations, exhibiting high novelty. Members who possess high knowledge interdependence but occupy low central position have limited access to core nodes. Independent research helps counteract the homogenization of the collaborative network caused by over-reliance on established knowledge, significantly driving the team's novelty. Regarding impact, teams characterized by high academic age diversity and high knowledge interdependence demonstrate a shared recognition for the team's technical knowledge base. When confronted with challenges of technological iteration, such teams benefit from a wide range of research perspectives, leading to a substantial improvement in innovation impact.
(4) In science-technology teams, academic inventors with low central position facilitate the generation of novel research outcomes. When both central position and collaborative tie strength are high, these strong collaborative foundations allow them to access accumulated expertise and frontier insights from influential partners, often leading to high novelty. Conversely, when central position is high but collaborative tie strength is low, academic age diversity can serve a compensatory role. Through the integration of diverse experiences and knowledge from different career stages, teams can achieve high levels of novelty via alternative mechanisms. Regarding impact, academic inventors with low collaborative tie strength and high academic age diversity are better positioned to minimize collaboration conflicts. By drawing on the varied resources and perspectives of members across academic generations, these teams continuously build and enhance their impact performance.

Implications

Through its findings, this study offers corresponding insights into building high performance research team. First, our findings demonstrate that low knowledge interdependence emerges as a key driver of novelty-oriented innovation, whereas high knowledge interdependence plays a central role in enhancing impact-related performance. Managers should hold a full understanding of their degree of knowledge interdependence, and to seek the "optimal performance balance" between novelty and impact. The "giants" in the phrase, "standing on the shoulders of giants", might significantly accelerate the progress in science and technology and enhance outcomes（Jiao et al., 2022）. Team managers can build a knowledge base of authoritative literature and systematically track the dynamics of leading scholars in the field, including their latest papers, speeches, collaborative projects, etc., integrate

this information into the team's knowledge system, and strengthen the sensitivity and integration of the knowledge system of the former "giants" within the team. To achieve high novelty, based on its own knowledge base and innovation needs, managers can adopts team member flow strategy and team knowledge management strategy to regularly recruit or mobilize new members with different knowledge coupling relationships to make better use of external knowledge and reduce the loss of knowledge value caused by excessive dependence on knowledge.

Second, this study demonstrates that the combined effects of network embeddedness on innovation performance show significant differences within different categories of research teams. Managers ought to select the appropriate performance improvement strategies based on the type of team. For scientific teams, encourage the formation of "junior - intermediate - senior" members with different qualifications to form a hierarchical research team, pay attention to balance the leadership of senior experts and the innovative vitality of young scholars, and promote the knowledge transfer and experience sharing of inter-generation cooperation through the form of mentor system or project cooperation, so as to accelerate the scientific research process and improve innovation performance. For technological teams, low knowledge interdependence can independently lead to high novelty innovations. Technological teams typically serve as R&D intensive teams. The output of novel ideas is the cumulative result of innovative knowledge search conducted by inventor teams (Beaudry et al., 2013). Team managers should pay attention to the introduction of core inventors(Li et al., 2013) to maintain the circulation and coordination of knowledge and information within the interconnected group of members, thus preventing "technological lock-in" and ensuring inventive innovations. For science-technology teams, a decision-making path characterized by low tie strength and high academic age diversity is conducive to teams obtaining both high novelty and high impact innovations. Managers should give academic inventors more autonomy in decision-making, such as setting regular meeting times to avoid information overload and to make the most of their strengths. Different academic backgrounds of team members should be considered when building teams, so as to enhance the team's ability to explore and acquire knowledge (Wu et al., 2024).

Third, the research results show that achieving high innovation performance requires the integration of multiple dimensions and factors, and the formation of high research team performance does not depend on a fixed combination of antecedent variables, and there are multiple feasible ways. The manager should clearly define the team's goals, comprehensively evaluate the team's current situation and advantages, and reasonably formulate the path of resource combination improvement. Faced with the challenge of path deviation, managers should adopt dynamic and adaptive exploratory thinking, establish a risk prevention and early warning mechanism, and timely discover alternative paths for equivalent innovation performance improvement factors. For example, when it is difficult for science and technology teams to maintain high cooperation intensity, scholars with different academic qualifications can be actively invited to join, that is, the diversity of academic age can be used as a substitute factor to help improve innovation performance. This

approach embodies the principle that different paths can lead to the same goal of enhancing innovation outcomes.

Limitations and future research

Despite all of the efforts, this study suffers from certain limitations. First, the analysis focuses exclusively on the pharmaceutical field, which may limit the generalizability of the findings. Future research could extend the scope to additional domains, comparing domain differences based on full domain data to draw more universally applicable conclusions. Second, the exploration of innovation performance pathways is grounded solely in the three dimensions of network embeddedness theory, which may not capture the full range of factors influencing research team performance. Future studies could consider additional variables, such as market influence, the sustained innovation capabilities of teams, levels of interdisciplinary collaboration, knowledge-sharing practices, and the influence of leadership.

Acknowledgments:

This work was supported by the National Natural Science Foundation of China (Grant No. 72174016 and 72474016).

References

Ardito, L., Natalicchio, A., Appio, F. P., & Petruzzelli, A. M. (2021). The role of scientific knowledge within inventing teams and the moderating effects of team internationalization and team experience: Empirical tests into the aerospace sector. Journal of Business Research, 128, 701-710.

Ashley, E. A. (2015). The precision medicine initiative: a new national effort. Jama, 313(21), 2119-2120.

Bai, H., Feng, F., Wang, J., & Wu, T. (2020). A combination prediction model of long-term ionospheric foF2 based on entropy weight method. Entropy, 22(4), 442.

Beaudry, C., & Schiffauerova, A. (2011). Impacts of collaboration and network indicators on patent quality: The case of Canadian nanotechnology innovation. European Management Journal, 29(5), 362-376.

Boxu, Y., Xingguang, L., & Kou, K. (2022). Research on the influence of network embeddedness on innovation performance: Evidence from China's listed firms. Journal of Innovation & Knowledge, 7(3), 100210.

Brass, D. J. (1984). Being in the right place: A structural analysis of individual influence in an organization. Administrative science quarterly, 518-539.

Breitzman, A. (2021). The relationship between web usage and citation statistics for electronics and information technology articles. Scientometrics, 126(3), 2085-2105.

Burt, R. S. (2018). Structural holes. In Social stratification (pp. 659-663). Routledge.

Cepoi, V., & Pandiloska Jurak, A. (2023). Measuring the relevance and impact of innovation and social forces for Transnational Value Chain's embeddedness in a region. Plos one, 18(10), e0291646.

Chien, S. Y., & Tsai, C. H. (2021). Entrepreneurial orientation, learning, and store performance of restaurant: The role of knowledge-based dynamic capabilities. Journal of Hospitality and Tourism Management, 46, 384-392.

Coleman, J. S. (1988). Social capital in the creation of human capital. American journal of sociology, 94, S95-S120.

Deng, J., Zhao, Y., Li, X., Wang, Y., & Zhou, Y. (2023). Network Embeddedness, Relationship Norms, and Cooperative Behavior: Analysis Based on Evolution of Construction Project Network. Journal of Construction Engineering and Management, 149(9), 04023070.

Dong, Y., Ma, H., Tang, J., & Wang, K. (2018). Collaboration diversity and scientific impact. arXiv preprint arXiv:1806.03694.

Granovetter, M. S. (1973). The strength of weak ties. American journal of sociology, 78(6), 1360-1380.

Guan, J., Zuo, K., Chen, K., & Yam, R. C. (2016). Does country-level R&D efficiency benefit from the collaboration network structure?. Research Policy, 45(4), 770-784.

Haeussler, C., & Sauermann, H. (2020). Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. Research Policy, 49(6), 103987.

Hao, X., Liang, Y., Yang, C., Wu, H., & Hao, Y. (2024). Can industrial digitalization promote regional green technology innovation?. Journal of Innovation & Knowledge, 9(1), 100463.

Harvey, S., & Berry, J. W. (2023). Toward a meta-theory of creativity forms: How novelty and usefulness shape creativity. Academy of Management Review, 48(3), 504-529.

Jiao, H., Wang, T., & Yang, J. (2022). Team structure and invention impact under high knowledge diversity: An empirical examination of computer workstation industry. Technovation, 114, 102449.

Kramoliš, J., & Šviráková, E. (2019). The influence of design on companies' increase in income, market share or brand value. Journal of Competitiveness.

Lee, R. P. (2010). Extending the environment–strategy–performance framework: The roles of multinational corporation network strength, market responsiveness, and product innovation. Journal of International Marketing, 18(4), 58-73.

Lee, Y. N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. Research policy, 44(3), 684-697.

Leydesdorff, L. (2018). Diversity and interdisciplinarity: how can one distinguish and recombine disparity, variety, and balance?. Scientometrics, 116, 2113-2121.

Li, Q., Maggitti, P. G., Smith, K. G., Tesluk, P. E., & Katila, R. (2013). Top management attention to innovation: The role of search selection and intensity in new product introductions. Academy of Management Journal, 56(3), 893-916.

Liu, J., Ding, K., Wang, F., Bu, Yi., & Maus, G. (2019). The structure and evolution ofscientific collaboration from the perspective of symbiosis. Malaysian journal of Library & Information Science, 24(1), 59–73.

Liu, J., Gong, X., Xu, S., & Huang, C. (2024). Understanding the relationship between team diversity and the innovative performance in research teams using decision tree algorithms: evidence from artificial intelligence. Scientometrics, 129(12), 7805-7831.

Liu, J., Guo, X., Xu, S., Bu, Y., Sugimoto, C. R., Larivière, V., ... & Zhou, H. (2024). Understanding super-partnerships in scientific collaboration: Evidence from the field of economics. Journal of the Association for Information Science and Technology, 75(6), 717-733.

Liu, Z., Xiang, B., Guo, W., Chen, Y., Guo, K., & Zheng, J. (2019). Overlapping community detection algorithm based on coarsening and local overlapping modularity. IEEE access, 7, 57943-57955.

Lyu, D., Gong, K., Ruan, X., Cheng, Y., & Li, J. (2021). Does research collaboration influence the “disruption” of articles? Evidence from neurosciences. Scientometrics, 126, 287-303.

Ma, B., & Zhang, J. (2022). Tie strength, organizational resilience and enterprise crisis management: An empirical study in pandemic time. International Journal of Disaster Risk Reduction, 81, 103240.
Mao, C., Xi, C., Du, R., Chen, W., Song, N., Qian, Y., & Tian, X. (2024). Characteristics of gut flora in children who go to bed early versus late. Scientific Reports, 14(1), 23256.
Mark, G. (1992). Problems of explanation in economic sociology. Networks and organizations: Structure, form, and action, 25-56.
Nordt, A., Raven, R., Malekpour, S., & Sharp, D. (2024). Actors, agency, and institutional contexts: Transition intermediation for low-carbon mobility transition. Environmental Science & Policy, 154, 103707.
Patel, P. C., & Jayaram, J. J. (2024). The impact of operating lease reporting rules on firm efficiency: the moderating role of supply chain network structure. International Journal of Production Research, 1-18.
Petersen, A. M. (2015). Quantifying the impact of weak, strong, and super ties in scientific careers. Proceedings of the National Academy of Sciences, 112(34), E4671-E4680.
Polanyi, K. (2002). The great transformation. Readings in economic sociology, 38-62.
Rishika, R., & Ramaprasad, J. (2019). The effects of asymmetric social ties, structural embeddedness, and tie strength on online content contribution behavior. Management Science, 65(7), 3398-3422.
Rowley, T., Behrens, D., & Krackhardt, D. (2000). Redundant governance structures: An analysis of structural and relational embeddedness in the steel and semiconductor industries. Strategic management journal, 21(3), 369-386.
Salazar, M. R., & Lant, T. K. (2018). Facilitating innovation in interdisciplinary teams: The role of leaders and integrative communication. Informing Science, 21, 157–178.
Schlattmann, S. (2016). Capturing the collaboration intensity of research institutions using social network analysis. Procedia Computer Science, 106, 25–31.
Schmidt, C. G., Yan, T., Wagner, S. M., & Lucianetti, L. (2022). Performance implications of knowledge inputs in inter-organisational new product development projects: the moderating roles of technology interdependence. International Journal of Production Research, 60(20), 6048-6071.
Schubert, T., & Tavassoli, S. (2020). Product innovation and educational diversity in top and middle management teams. Academy of Management Journal, 63(1), 272-294.
Schweitzer, F., Garas, A., Tomasello, M. V., Vaccario, G., & Verginer, L. (2022). The role of network embeddedness on the selection of collaboration partners: An agent-based model with empirical validation. Advances in Complex Systems, 25(02n03), 2250003.
Sheng, J., Liang, B., Wang, L., & Wang, X. (2024). A study on citation impact with age diversity among disciplines. Physica A: Statistical Mechanics and its Applications, 653, 130096.
Singh, R. P. (2000). Entrepreneurial opportunity recognition through social networks. Psychology Press.
Song, H., Chen, R., Yang, X., & Hou, J. (2024). How Does the Innovation Openness of China's Sci-Tech Innovation Enterprises Support Innovation Quality: The Mediation Role of Structural Embeddedness. Mathematics, 12(19), 3034.
Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. Academy of management journal, 44(2), 316-325.
Stratton, C., Christensen, A., Jordan, C., Salvatore, B. A., & Mahdavian, E. (2024). An interdisciplinary course on computer-aided drug discovery to broaden student

participation in original scientific research. Biochemistry and Molecular Biology Education.
Tian, Q., Li, G., & Xu, R. (2021, December). Research on the Impact of Network Embeddedness on Enterprise Innovation Performance--Based on the Mediating Role of Business Model Innovation and the Moderating Role of Competition Intensity. In 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1047-1051). IEEE.
Tortoriello, M. (2015). The social underpinnings of absorptive capacity: The moderating effects of structural holes on innovation generation based on external knowledge. Strategic Management Journal, 36(4), 586-597.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. Science, 342(6157), 468-472.
Wang, J. (2016). Knowledge creation in collaboration networks: Effects of tie configuration. Research policy, 45(1), 68-80.
Wang, Y., Zhang, J., Yan, Y., & Guan, J. (2024). The bidirectional causality of tie stability and innovation performance. Research Policy, 53(10), 105102.
Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.
Wu, K., Xie, Z., & Du, J. T. (2024). Does science disrupt technology? Examining science intensity, novelty, and recency through patent-paper citations in the pharmaceutical field. Scientometrics, 129(9), 5469-5491.
Xie, X., Fang, L., & Zeng, S. (2016). Collaborative innovation network and knowledge transfer performance: A fsQCA approach. Journal of business research, 69(11), 5210-5215.
Xu, S., Li, L., & An, X. (2023). Do academic inventors have diverse interests?. Scientometrics, 128(2), 1023-1053.
Xu, S., Li, L., An, X., Hao, L., & Yang, G. (2021). An approach for detecting the commonality and specialty between scientific publications and patents. Scientometrics, 126, 7445-7475.
Yan, Y., Zhang, J., & Guan, J. (2019). Network embeddedness and innovation: Evidence from the alternative energy field. IEEE Transactions on Engineering Management, 67(3), 769-782.
Yang, Y., She, Y., Hong, J., & Gan, Q. (2021). The centrality and innovation performance of the quantum high-level innovation team: the moderating effect of structural holes. Technology Analysis & Strategic Management, 33(11), 1332-1346.
Yoo, H. S., Jung, Y. L., Lee, J. Y., & Lee, C. (2024). The interaction of inter-organizational diversity and team size, and the scientific impact of papers. Information Processing & Management, 61(6), 103851.
Zaheer, A., Gözübüyük, R., & Milanov, H. (2010). It's the connections: The network perspective in interorganizational research. Academy of management perspectives, 24(1), 62-77.
Zhang, H., Tan, X., Liu, C., & Chen, M. (2023). Do Team Boundary-Spanning Activities Affect Innovation Performance?. Sustainability, 15(13), 10605.
Zhang, L., Qiu, H., Chen, J., Zhou, W., & Li, H. (2023). How Do Heterogeneous Networks Affect a Firm's Innovation Performance? A Research Analysis Based on Clustering and Classification. Entropy, 25(11), 1560.
Zhang, N., Cheng, L., Sun, C., & Van Looy, B. (2023). The role of inter-and intra-organisational networks in innovation: towards requisite variety. Scientometrics, 128(7), 4117-4136.

Zhang, P., Han, S., & Cai, Z. (2021). Publication Month Bias Evolution Patterns of Highly Cited Papers in Different Disciplines. In 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1062-1066). IEEE.

Zhao, S. X., Chen, D. Z., Huang, M. H., & Chang, Y. W. (2019). Potential value of patents with provisional applications: an assessment of bibliometric approach. IEEE Transactions on Engineering Management, 69(6), 2497-2516.

Zhao, Y., Wang, Y., Zhang, H., Kim, D., Lu, C., Zhu, Y., & Zhang, C. (2024). Do more heads imply better performance? An empirical study of team thought leaders' impact on scientific team performance. Information Processing & Management, 61(4), 103757.

Zhou, J., & Cen, W. (2024). Digital Entrepreneurial Ecosystem Embeddedness, Knowledge Dynamic Capabilities, and User Entrepreneurial Opportunity Development in China: The Moderating Role of Entrepreneurial Learning. Sustainability, 16(11), 4343.

Zou, B., Wang, Y., Kwoh, C. K., & Cen, Y. (2023). Directed collaboration patterns in funded teams: A perspective of knowledge flow. Information Processing & Management, 60(2), 103237.

https://doi.org/10.51408/issi2025_058

Research-Policy Alignment in AI: A Bibliometric Study of the EU AI Act

Cristian Mejia

mejia.cristian@ifi.u-tokyo.ac.jp
Institute for Future Initiatives, University of Tokyo, Tokyo (Japan)

Abstract

The rapid advancement of artificial intelligence (AI) necessitates understanding how academic research aligns with emerging regulatory frameworks. This study employs topic modeling to examine the relationship between library and information science research and AI policy priorities. We analyzed 2,795 academic publications on AI in library science and 1,005 statements from the European Union's AI Act, identifying 56 research clusters and 33 regulatory topics, respectively. Using semantic similarity measures, we mapped thematic alignments between research and policy domains. Results reveal strong concordance in areas such as governance frameworks and risk management, while highlighting gaps in regulatory implementation research and domain-specific applications. Notable mismatches include limited academic engagement with regulatory bodies and oversight mechanisms, contrasting with substantial research focus on cultural heritage and medical applications that lack direct regulatory correspondence. This study contributes a systematic methodology for evaluating research-policy alignment in emerging technologies, building on established bibliometric approaches for assessing research impact on policy. Our findings suggest the need for enhanced dialogue between researchers and policymakers while demonstrating how academic inquiry extends beyond immediate regulatory concerns.

Introduction

The unprecedented advancement of artificial intelligence (AI) technologies has prompted governments worldwide to develop comprehensive regulatory frameworks, exemplified by landmark legislation such as the European Union's AI Act (European Parliament, 2024). As researchers in information and library science, we regularly contribute to the AI knowledge base through studies on implementation, governance, ethics, and technological applications. However, there remains a critical gap in understanding whether our collective research priorities align with the aspects of AI that policymakers seek to regulate. This alignment—or potential mismatch—between academic research focus and policy concerns carries significant implications for both the effectiveness of evidence-based policymaking and the societal impact of our research. To address this knowledge gap, we propose a systematic bibliometric approach comparing research trends in library and information science with areas of interest in policy documents, providing an objective assessment of the concordance between academic interests and regulatory priorities in the rapidly evolving AI landscape.

The relationship between research and policymaking has been a subject of longstanding academic interest, traditionally examined through qualitative approaches that analyze how research findings influence policy decisions and how policy priorities shape research agendas. Ritter and Lancaster (2013) demonstrated this through a case study of drug policy, highlighting that assessing research

influence requires examining multiple channels, including direct citations in policy documents, utilization within policy processes, and dissemination through media coverage. This multi-dimensional approach acknowledges that research impact on policy extends beyond simple citation metrics and involves complex interactions between researchers, policymakers, and other stakeholders.

As the field evolved, researchers developed more systematic and quantitative methods to assess the research-policy relationship. Van Leeuwen et al. (2003) pioneered work in bibliometric approaches to evaluate research excellence and its influence on science policy, shifting from average-based impact metrics toward indicators that better reflect top-performing research. This methodological evolution was further exemplified by Debackere and Glanzel (2004), who demonstrated how bibliometric data could support major funding allocation decisions, highlighting the practical application of systematic research evaluation in policy contexts.

A significant advancement in this field has been the development of specialized databases and tools for tracking policy impact. The Overton database represents a major milestone, providing comprehensive indexing of policy documents and their academic citations (Szomszor & Adie, 2022). This development has enabled more sophisticated analyses of how research influences policy across different disciplines and jurisdictions. However, as Newson et al. (2018) revealed in their study of obesity policy documents, citation-based approaches have limitations – policy documents don't always explicitly cite their academic sources, and when they do, these citations may not accurately reflect the actual influence of research on policy development.

To address these limitations, researchers have explored innovative text-based methods to identify connections between different knowledge domains. Ittipanuvat et al. (2014) employed Literature-Based Discovery (LBD) to uncover linkages between technological developments and social issues, demonstrating how text analysis can reveal previously hidden connections between research and societal needs. Similarly, Takano and Kajikawa (2019) utilized text similarity measures to identify commercialization opportunities by comparing academic papers with patents. These approaches show how computational text analysis can uncover implicit relationships between research outputs and their practical applications, even when explicit citations are absent. Such methodologies offer promising alternatives for understanding the complex relationship between academic research and policy development, particularly in rapidly evolving fields where traditional citation metrics might lag behind the pace of innovation.

These methodological approaches for analyzing research-policy relationships become particularly relevant in rapidly evolving technological domains where the need for evidence-based policymaking is crucial. Artificial intelligence represents one such domain, where the acceleration of technological capabilities has prompted unprecedented policy responses worldwide. In the past few years, we witnessed significant momentum in AI governance initiatives across different jurisdictions and international bodies. The G7 Hiroshima AI Process established the world's first international framework for AI governance (G7 Leaders, 2023), while the United Nations emphasized the need for AI regulation based on the UN Charter and Universal Declaration of Human Rights (Guterres, 2023). Organizations like

UNESCO have also contributed through their Recommendation on the Ethics of AI, offering the first global, rights-based framework for AI policy (UNESCO, 2022).

However, the European Union's AI Act, which came into force in August 2024, represents a watershed moment in AI regulation. Unlike previous policy instruments that primarily focused on ethical principles or voluntary guidelines, the EU AI Act establishes a comprehensive and legally binding regulatory framework. The Act introduces a sophisticated risk-based approach, categorizing AI applications into unacceptable, high, limited, and minimal risk levels, while also addressing the emerging challenges of general-purpose AI systems (European Parliament, 2024). Its extraterritorial scope means it affects AI providers worldwide who serve EU users, similar to the impact of GDPR on data protection practices globally (European Union, 2016).

In this context, examining how library and information science research aligns with the EU AI Act's regulatory framework becomes particularly valuable. Our field's research spans multiple dimensions of AI implementation, from technical applications in information retrieval and digital collections to broader considerations of ethics, governance, and user impact. By employing bibliometric and text analysis methods to compare research clusters with policy document clusters, we can identify areas where academic research effectively informs policy decisions and where potential gaps might exist. This systematic analysis can guide future research directions, ensure our field's relevance to policy development, and potentially reveal unique insights from our discipline that could inform future AI governance frameworks. Moreover, our methodological approach offers a replicable framework for assessing research-policy alignment in other rapidly evolving technological domains.

Data and Methods

Our analysis draws on two distinct datasets: academic publications indexed in Web of Science (WoS) and the European Union's Artificial Intelligence Act. We selected Web of Science Core Collection as our primary bibliometric data source due to its comprehensive coverage of high-quality academic literature and standardized citation tracking. WoS's detailed metadata ensure reliable bibliometric analysis, while its consistent categorization system enables precise field-specific queries.

Using the search query TS=("artificial intelligence") AND WC=("information science library science"), we extracted all document types across all available publication years. This search strategy captured articles *explicitly* acknowledging a focus on AI within the specific context of library and information science, yielding 2,795 records as of January 15, 2025. By not restricting document types or publication years, we ensured comprehensive coverage of how the field has engaged with AI-related topics over time.

For our policy analysis, we focused on the European Union's AI Act, downloaded in English from the official EU website in HTML format. The Act is structured into 12 main chapters plus a 13th chapter dedicated to amendments. While the complete legislation includes additional annexes and recitals, we rely on the main chapters to focus on the core regulatory provisions. To enable detailed content analysis, we

decomposed the Act into individual statements, treating each numerical subdivision within articles as a distinct unit of analysis. This granular approach resulted in 1,005 unique statements, providing a detailed representation of the Act's regulatory scope and requirements.

Our analytical framework employs topic modelling (Blei, 2012) to identify thematic structures within both academic publications and policy statements. For academic records, we preprocessed the data by concatenating titles and abstracts for each publication. Similarly, we prepared the policy statements by removing leading numerals while preserving the complete textual content of each regulatory provision. The topic modeling process utilized BERTopic (Grootendorst, 2022), a state-of-the-art library that leverages BERT's contextual embeddings to generate more semantically coherent topics compared to traditional approaches like LDA. To determine the optimal number of topics for each dataset, we conducted an iterative process testing configurations ranging from 10 to 200 topics, selecting the solution that maximized the coherence metric (Farea et al., 2024). This approach resulted in 56 topics for the academic dataset and 33 topics for the policy statements.

Each document was then assigned to its most probable topic, effectively creating distinct clusters within both datasets. For academic clusters, we calculated additional metrics including average publication year and mean citation count, providing temporal and impact dimensions to our analysis. We manually labeled each cluster based on careful examination of its constituent documents, considering frequent terms, representative papers, and thematic coherence.

To identify alignments between research priorities and policy concerns, we computed cosine similarity scores between the topic vectors of academic and policy clusters. This similarity metric captures semantic overlap between clusters, with higher scores indicating stronger thematic alignment. Cosine similarity is particularly suitable for this comparison as it normalizes for differences in document length and term frequency distributions between academic and policy texts.

This methodological framework enables systematic comparison between research focus areas and regulatory priorities, revealing both convergences and potential gaps between academic inquiry and policy development in the domain of artificial intelligence within library and information science.

Results and Discussion

Our analysis identified 56 distinct research topics from 2,795 academic publications on AI in library science, while the 1,005 statements extracted from the EU AI Act clustered into 33 regulatory topics. These two topic landscapes represent the research interests of academics and the regulatory priorities of policymakers, respectively. The results of the academic landscape analysis are presented in Figure 1 and Table 1.

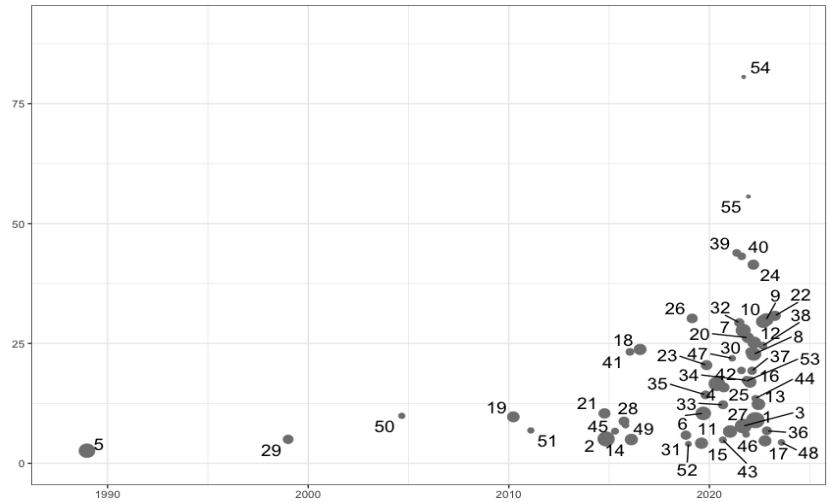


Figure 1. Temporal distribution and impact of AI research topics in library science. Each point represents a research cluster, with its position determined by average publication year (x-axis) and average citation count (y-axis). Numbers correspond to cluster IDs.

The visualization reveals the evolution of AI research within library science over the past three decades. Early research in the 1990s centered on fundamental information retrieval systems, as represented by cluster 5. The field has since undergone significant transformation, with recent research focusing on emerging technologies such as blockchain integration (cluster 55) and applications of generative AI (clusters 22, 36).

Table 1. Summary of selected research clusters on AI in library science, including the five most recent, most cited, and largest by number of documents.

Id	*Cluster name*	*Docs.*	*Ave. Year*	*Ave. Cites.*
1	Strategic Implementation of AI Technologies in Library Service Innovation	143	2,022.3	9.0
2	AI Integration and Digital Transformation in Information Management Systems	120	2,014.9	5.1
3	AI-Powered Content Analysis and Generation in Digital Media	113	2,021.7	7.8
4	Trust and Governance Frameworks for Healthcare AI Implementation	110	2,020.4	16.6
5	Applications of Artificial Intelligence in Information Retrieval Systems	110	1,989.0	2.6
9	Technology Acceptance Models and User Adoption Factors in AI-Enabled Systems	82	2,022.8	30.0
17	AI-Enhanced Peer Review Systems in Academic Publishing	61	2,022.8	4.7
22	AI-Assisted Knowledge Construction in Academic Research and Writing	49	2,023.3	30.8
24	AI Implementation Frameworks and Challenges in Organizational Systems	48	2,022.2	41.4

36	Chatbot Implementation in Libraries	36	2,022.9	6.8
39	AI-Driven Information Management Solutions During the COVID-19 Crisis	28	2,021.4	43.9
40	AI-Driven Marketing Analytics and Customer Segmentation Systems	28	2,021.6	43.2
48	Digital Literacy Evolution in the AI Era	22	2,023.6	4.4
54	Big Data Analytics Applications in Organizational Decision Support Systems	17	2,021.7	80.6
55	Blockchain Integration in Information Systems	17	2,021.9	55.6

The citation patterns reveal varying levels of scholarly impact across research topics. Big data analytics (cluster 54), blockchain applications (cluster 55), and COVID-19 related research (cluster 39) have garnered recent attention, each averaging over 40 citations per paper. Implementation frameworks (cluster 24), trust dynamics (cluster 38), and organizational impact studies (cluster 10) have also demonstrated substantial influence with moderate citation rates.

Perhaps most notably, we see a marked concentration of research clusters in the 2020-2024 period, indicating an acceleration of AI-related research within library science. This temporal clustering coincides with the development and implementation of the EU AI Act, suggesting a potential alignment between academic research priorities and emerging regulatory frameworks. This synchronicity provides a valuable foundation for examining the relationship between research focus areas and regulatory priorities.

The analysis of the EU AI Act yielded 33 distinct clusters that reflect the regulatory framework's key priorities as seen in Table 2. These clusters broadly align into several core themes. The foundational elements of the Act are represented in clusters focusing on governance structures, including the establishment of the AI Office, Scientific Panel, and oversight mechanisms (clusters 4, 16, and 22). A significant portion of clusters addresses specific technical and operational requirements, such as conformity assessment procedures (cluster 5), data processing protocols (cluster 12), and logging requirements (cluster 33).

Table 2. Summary of clusters from the EU AI Act. The top largest clusters are shown.

Id	*Cluster name*	*Statements*
1	Notified Bodies	56
2	Market Surveillance and Law Enforcement Authority Framework	52
3	SME and Start-up Support Mechanisms	49
4	Governance and Advisory Bodies Structure	48
5	Conformity Assessment and Documentation Requirements	46
6	General-Purpose AI Models Classification and Risk Management	46
7	Technical Requirements for High-Risk AI Systems	45
8	Market Placement and Compliance Requirements	43
9	Stakeholder Obligations and Responsibilities	42
10	Risk Assessment and Harm Classification	40

Risk management emerges as a central theme, with dedicated clusters covering risk assessment methodologies (cluster 10), market surveillance (clusters 2 and 11), and incident reporting frameworks (cluster 32). The Act's emphasis on documentation and transparency is reflected in clusters focusing on provider obligations (cluster 30), compliance documentation (cluster 26), and certificate management (cluster 29).
Notably, several clusters specifically address emerging technologies and their regulatory implications, particularly in the context of biometric systems (cluster 18) and general-purpose AI models (cluster 6). The Act also maintains focus on practical implementation through clusters dedicated to SME support mechanisms (cluster 3), market access requirements (cluster 19), and administrative procedures (cluster 28). This clustering reveals the Act's comprehensive approach to AI regulation, balancing high-level governance principles with specific technical requirements and practical implementation considerations. The distribution of topics suggests a regulatory framework that aims to be both thorough in its coverage and pragmatic in its application.

Linkage between academic research and policy

After analyzing the individual topic landscapes, we examined the thematic alignment between academic research clusters and regulatory topics through semantic similarity analysis. When two clusters from different datasets show high similarity, this indicates that the academic research focus substantively overlaps with regulatory priorities in that area. Academic research and policy documents are written in different styles and use different vocabulary; thus, high absolute similarity is not expected. Therefore, we define as similar pairs those beyond the third quartile across all possible connections (i.e., >0.46), suggesting relative strong thematic concordance between the research focus and policy considerations. The similar pairs are shown in Figure 2.

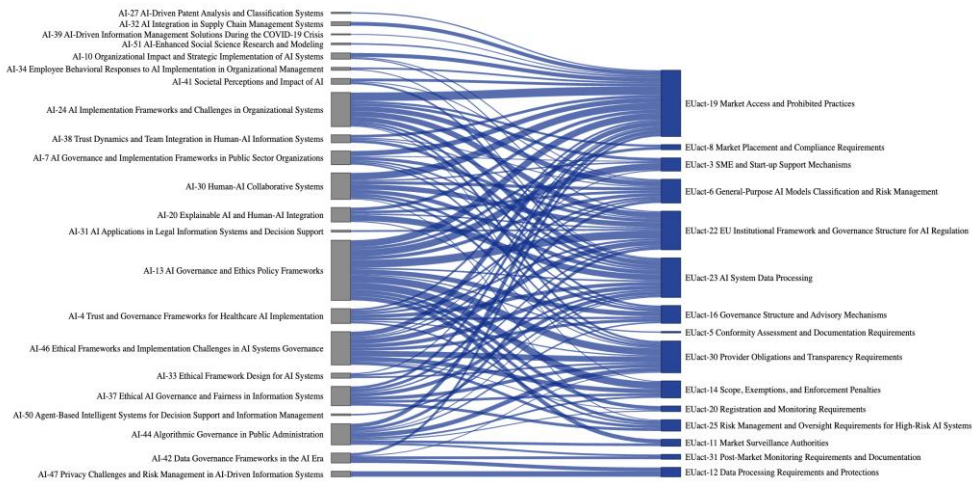


Figure 2. Semantic linkages between AI research topics in library science (left) and EU AI Act regulatory clusters (right). The width of connecting lines represents the strength of thematic similarity between clusters. Only connections with similarity scores above 0.46 are shown.

The Sankey diagram reveals several notable alignments between research and policy domains. A particularly strong connection exists between academic research on "AI Governance and Ethics Policy Frameworks" (cluster 13) and the regulatory focus on "Market Surveillance and Law Enforcement Authority Framework" (cluster 2). This alignment suggests that academic research has been actively engaging with governance challenges that policymakers consider crucial.

Another significant match appears between research on "Trust Dynamics and Team Integration in Human-AI Information Systems" (cluster 38) and the regulatory cluster on "Risk Management and Oversight Requirements for High-Risk AI Systems" (cluster 25). This pairing indicates that academic investigations into human-AI interaction and trust align well with regulatory concerns about risk management in high-stakes AI applications.

However, the visualization also reveals areas where academic research and regulatory focus may not fully align, as evidenced by clusters with few or no strong connections. This pattern suggests opportunities for future research to address emerging regulatory priorities.

Research gaps in relation to regulatory clusters

The analysis reveals notable gaps between academic research priorities and certain regulatory focuses. Particularly striking is the limited academic engagement with regulatory bodies and administrative frameworks, which are central to clusters 1 ("Notified Bodies") and 4 ("Governance and Advisory Bodies Structure") of the EU AI Act. While these clusters detail the operational mechanics of AI oversight - including the roles of notified bodies in conformity assessment and the structure of advisory forums - our bibliometric analysis shows minimal research addressing these institutional aspects within library and information science.

This mismatch likely stems from the traditionally technical and user-focused nature of library science research, which has emphasized practical implementations and user interactions with AI systems rather than regulatory mechanisms. However, this gap presents valuable research opportunities. Future studies could examine how information institutions interact with regulatory bodies, how conformity assessments impact information services, and how library and information science expertise could inform the development of AI governance structures. Additionally, research investigating the role of libraries and information centers as potential intermediaries in the regulatory framework could provide valuable insights for both policymakers and practitioners.

Regulatory gaps in relation to research clusters

The analysis also reveals areas where academic research has developed substantial focus that is not directly reflected in the regulatory framework. For instance, clusters AI-15 ("AI-Enabled Digital Collection Management in Cultural Institutions") and AI-16 ("Clinical Applications of AI in Medical Diagnosis and Prognosis") represent significant research streams with limited corresponding regulatory attention in the EU AI Act.

In the case of digital collection management (AI-15), this mismatch likely reflects the specialized nature of cultural heritage applications, which may not warrant specific regulatory attention despite their importance to the library and information science community. The research in this area focuses on practical implementations and professional practices that fall under broader regulatory categories rather than requiring dedicated regulatory frameworks.

Similarly, while medical AI applications (AI-16) represent a crucial research area within our field, their regulation is primarily addressed through specialized healthcare frameworks and medical device regulations rather than the general-purpose AI Act. This suggests that some domain-specific AI applications, though important in academic research, may be better governed through sector-specific regulatory instruments rather than general AI legislation.

Conclusion

In this study, we employed bibliometric analysis and topic modeling to examine the alignment between academic research in library science and AI policy priorities as reflected in the EU AI Act. By analyzing 2,795 academic publications and 1,005 policy statements, we identified 56 research clusters and 33 regulatory topics, enabling a systematic comparison of thematic focus areas through semantic similarity measures.

Our findings resonate with previous research on evidence-based policymaking. As Van Leeuwen et al. (2003) emphasized the need for sophisticated metrics to evaluate research excellence, our analysis provides a quantitative framework for assessing research-policy alignment. Our proposal also highlights the role of bibliometrics in providing new angles that may facilitate the work of policymakers (Kajikawa, 2022). The identification of both matches and mismatches in our results supports Ritter and Lancaster's (2013) assertion that research influence on policy operates through multiple channels and complex interactions.

The study reveals both encouraging alignments and notable gaps between academic research and regulatory priorities. While we found strong concordance in areas such as governance frameworks and risk management, significant disparities emerged in others. These findings underscore the need for enhanced dialogue between researchers and policymakers in shaping AI governance within information environments. Similarly, the presence of research clusters with limited regulatory correspondence demonstrates how academic inquiry naturally extends beyond immediate regulatory concerns to address domain-specific challenges.

This research makes a novel contribution by providing a systematic, quantitative methodology for evaluating the relationship between research priorities and regulatory frameworks in rapidly evolving technological domains. Our approach offers a replicable framework for assessing research-policy alignment that could be applied to other emerging technologies and regulatory contexts.

Several limitations and opportunities for future research exist. First, our analysis focuses solely on the EU AI Act; incorporating other regulatory frameworks such as the Council of Europe AI Treaty and White House Executive Orders would provide a more comprehensive view. Second, expanding the analysis beyond library science

or broadening the AI-related search terms could offer wider perspectives on research-policy alignment. Finally, analyzing research clusters by country of origin could reveal geographical variations in research priorities and their relationship to national policy approaches.

Acknowledgments

This research is supported by a Grant in Aid for Early Career Scholars from the Japan Society for the Promotion of Science [24K16438].

References

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Debackere, K., & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics, 59*(2), Article 2. https://doi.org/10.1023/B:SCIE.0000018532.70146.02

European Parliament. (2024). *Regulation 2024/1689. Artificial Intelligence Act* (OJ L, 2024/1689, 12.7.2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj#document1

European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)* (32016R0679; Official Journal of the European Union). https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

Farea, A., Tripathi, S., Glazko, G., & Emmert-Streib, F. (2024). Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research. *Engineering Applications of Artificial Intelligence, 136*, 108877. https://doi.org/10.1016/j.engappai.2024.108877

G7 Leaders. (2023). *Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI system*. G7 Hiroshima Summit 2023.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794; Version 1). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Guterres, A. (2023, July 18). *UN Secretary-General's remarks to the Security Council on Artificial Intelligence*. https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence

Ittipanuvat, V., Fujita, K., Sakata, I., & Kajikawa, Y. (2014). Finding linkage between technology and social issue: A Literature Based Discovery approach. *Journal of Engineering and Technology Management, 32*, 160–184. https://doi.org/10.1016/j.jengtecman.2013.05.006

Kajikawa, Y. (2022). Reframing evidence in evidence-based policy making and role of bibliometrics: Toward transdisciplinary scientometric research. *Scientometrics, 127*(9), Article 9. https://doi.org/10.1007/s11192-022-04325-6

Newson, R., Rychetnik, L., King, L., Milat, A., & Bauman, A. (2018). Does citation matter? Research citation in policy documents as an indicator of research impact – an Australian obesity policy case-study. *Health Research Policy and Systems, 16*(1), 55. https://doi.org/10.1186/s12961-018-0326-9

Ritter, A., & Lancaster, K. (2013). Measuring research influence on drug policy: A case example of two epidemiological monitoring systems. *International Journal of Drug Policy*, *24*(1), 30–37. https://doi.org/10.1016/j.drugpo.2012.02.005

Szomszor, M., & Adie, E. (2022). Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, *3*(3), 624–650. https://doi.org/10.1162/qss_a_00204

Takano, Y., & Kajikawa, Y. (2019). Extracting commercialization opportunities of the Internet of Things: Measuring text similarity between papers and patents. *Technological Forecasting and Social Change*, *138*, 45–68. https://doi.org/10.1016/j.techfore.2018.08.008

UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence* (Programme and Meeting Document SHS/BIO/PI/2021/1). https://unesdoc.unesco.org/ark:/48223/pf0000381137

Van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & Van Raan, A. F. J. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, *57*(2), Article 2. https://doi.org/10.1023/A:1024141819302

https://doi.org/10.51408/issi2025_064

Retracted Citations and Self-citations in Retracted Publications: A Comparative Study of Plagiarism and Fake Peer Review

Kiran Sharma[1], Parul Khurana[2]

[1]*kiran.sharma@bmu.edu.in*
School of Engineering & Technology, BML Munjal University, Gurugram, Haryana-122413 (India)
Center for Advanced Data and Computational Science, BML Munjal University, Gurugram, Haryana-122413 (India)

[2]*parul.khurana@lpu.co.in*
School of Computer Applications, Lovely Professional University, Phagwara, Punjab-144411 (India)

Abstract

Retracted citations remain a significant concern in academia as they perpetuate misinformation and compromise the integrity of scientific literature despite their invalidation. To analyze the impact of retracted citations, we focused on two retraction categories: plagiarism and fake peer review. The data set was sourced from Scopus and the reasons for the retraction were mapped using the Retraction Watch database. The retraction trend shows a steady average growth in plagiarism cases of 1.2 times, while the fake peer review exhibits a fluctuating pattern with an average growth of 5.5 times. Although fewer papers are retracted in the plagiarism category compared to fake peer reviews, plagiarism-related papers receive 2.5 times more citations. Furthermore, the total number of retracted citations for plagiarized papers is 1.8 times higher than that for fake peer review papers. Within the plagiarism category, 46% of the retracted citations are due to plagiarism, while 53.6% of the retracted citations in the fake peer review category are attributed to the fake peer review. The results also suggest that fake peer review cases are identified and retracted more rapidly than plagiarism cases. Finally, self-citations constitute a small percentage of citations to retracted papers but are notably higher among citations that are later retracted in both the categories.

Introduction

Retracted citations refer to citations made to academic papers that have been officially retracted by publishers or journals due to issues such as errors, misconduct, plagiarism, falsified data, or ethical violations. Despite being retracted, these papers often continue to be cited in new research, sometimes without acknowledgment of their retracted status (Gray et al., 2019; da Silva, 2020; Silva and Bornemann-Cimenti, 2016). The issue of retracted citations poses a serious challenge to the academic community, as retracted papers often continue to be cited despite their invalidation. This practice can spread misinformation and undermine the credibility and integrity of the scientific literature.

The number of citations for retracted articles has increased over time, with a constant increase in the percentage of acknowledging their retraction (Heibi and Peroni, 2021; Sharma, 2024). Most of the retracted articles, particularly those published in Nature, Science, and Cell, continue to be cited even after their retraction (Wang and Su, 2022). Tang (2023) study also highlighted that post-retraction citations in the top

ranked journals, Nature and Science, account for 47.7% and 40.9% of total citations, respectively, with factors such as misconduct, validity issues, and background citation noise contributing to these retractions.

Post-retraction citations are an avoidable phenomenon. Although, retraction decreased the frequency of citation by about 60%, compared to non-retracted papers, but retracted papers often live on (Kühberger et al., 2022). Previous research on retracted articles has revealed that, despite being flagged, such studies are still frequently cited as valid across various disciplines (Bar-Ilan and Halevi, 2017; Sharma, 2021).

In the study by Cassai et al. (2022) in anesthesiology and intensive care medicine, they examined that 46% of the articles retracted were cited at least once after retraction, and many authors were unaware of the retraction. Bolboacă et al. (2019) investigates the trends and citation patterns that occur after retraction of articles in the field of radiology imaging diagnostics. Post-retraction citations in radiology imaging diagnostic methods are higher than before retraction in 30 out of 54 cases, plagiarism being the most common reason for retraction (31%). The persistence of post-retraction citations in radiology-imaging diagnostic methods, as well as in other medical fields like radiation oncology, points to a systemic issue in academic publishing.

Retracted biomedical research papers continue to be cited at relatively high rates, despite the retraction process (Hagberg, 2020). Hamilton (2019) also quantified the number and explored the nature of the citations of articles retracted in the radiation oncology literature that occur after publication of the retraction note. The study found that 92% of the 358 post-retraction citations examined referenced retracted articles as legitimate work. The results of the study emphasize the need for investigators to adhere to good research practices to mitigate the influence and propagation of flawed and unethical research. Schneider et al. (2020) also presented a case study of long-term post-retraction citation to falsified clinical trial data. They investigated that even 11 years after its retraction, the paper is still being cited positively and uncritically to support a medical nutrition intervention, with no acknowledgment of its 2008 retraction for data falsification.

In addition, Palla et al. (2023) studied that the number of articles retracted by Indian researchers increased from 2001 to 2020. The main reason for the retraction was duplication and plagiarism. They analyzed that 90% of the articles retracted by Indian researchers were cited even after the retraction process, with a total decline of 8% in citations after the retraction process. The protocol proposed by Heibi and Peroni (2022) can be used as a comprehensive framework to analyze the citation patterns of retracted articles. This is due to the importance of increasing awareness and better management of the retraction information. Understanding such patterns can therefore help mitigate the impact of retracted articles on the scientific literature and ensure academic research integrity.

Research Gap

Earlier studies have focused mainly on the increase in citations of retracted publications in various fields. Koçyiğit et al. (2023) analyzed retracted articles in the

medical literature due to ethical issues. Qi et al. (2016) studied the retraction due to fake peer reviews, where publishing journals and authors are concerned. Kamali et al. (2020) studied Iran-associated scientific papers retracted for duplication, plagiarism, and fake peer reviews, calling for immediate intervention and education in ethical research. Wang and Chen (2025) studied retractions due to honest errors in team size. On the other hand, Rivera (2018) highlighted that inappropriate authorship and fake peer review are real evils that contribute to lower quality publications. Bell et al. (2022) also highlighted that fake peer reviews in scholarly publications are a growing concern, highlighting the need to distinguish genuine reviews and to defend the boundaries between science and society. All of these papers examined various reasons for retractions in the context of the growing number of retracted publications. However, none focused on retracted citations, their subsequent reasons, or self-citations. This study addresses this gap by specifically analyzing the retraction categories of plagiarism and fake peer review, their associated retracted citations, and the reasons behind these retractions. Additionally, it delves deeper into the self-citations reported within retracted citations in both categories.

Research Objectives

The study seeks to analyze and investigate the following objectives within the categories of plagiarism and fake peer review retractions:

1. Trends in publication and retraction for both categories.
2. Distribution of retracted citations across both categories and the reason for retractions.
3. Distribution of self-citations within both categories.

Research Questions

The following research questions are designed to guide the investigation of the objectives related to plagiarism and fake peer review retractions:

- **R1:** How have the retraction rates evolved over time in both categories?
- **R2:** What is the average time of retraction (in years) for plagiarism and fake peer review?
- **R3:** What is the distribution of retracted citations in plagiarism and fake peer review retractions?
- **R4:** What share of retracted citations falls under the same retracted category?
- **R5:** How do self-citations contribute to the total number of citations in plagiarized and fake peer review retracted articles?
- **R6:** Are self-citations more prevalent in one category (plagiarism or fake peer review retraction) than in the other?

Methodology

Data Description

The study utilized Scopus-sourced scholar publishing data, downloaded on 7 December 2024, comprising a total of 33,188 publications with document type retracted. The Scopus query to extract the retracted publication was: "DOCTYPE (tb)". To ensure accurate linkage with retraction records, the dataset was filtered to include only documents with a DOI Khurana et al. (2022), resulting in 32,861 entries. These DOIs were then matched with the Retraction Watch database (https://www.crossref.org/blog/news-crossref-and-retraction-watch/) to identify documents flagged for retraction. This mapping process was successful for 26,908 documents. Subsequently, a filtration step was applied to isolate cases where the nature of the retraction explicitly indicated "Retraction", which produced a data set of 26,528 documents for analysis. Figure 1 shows the description of the data.

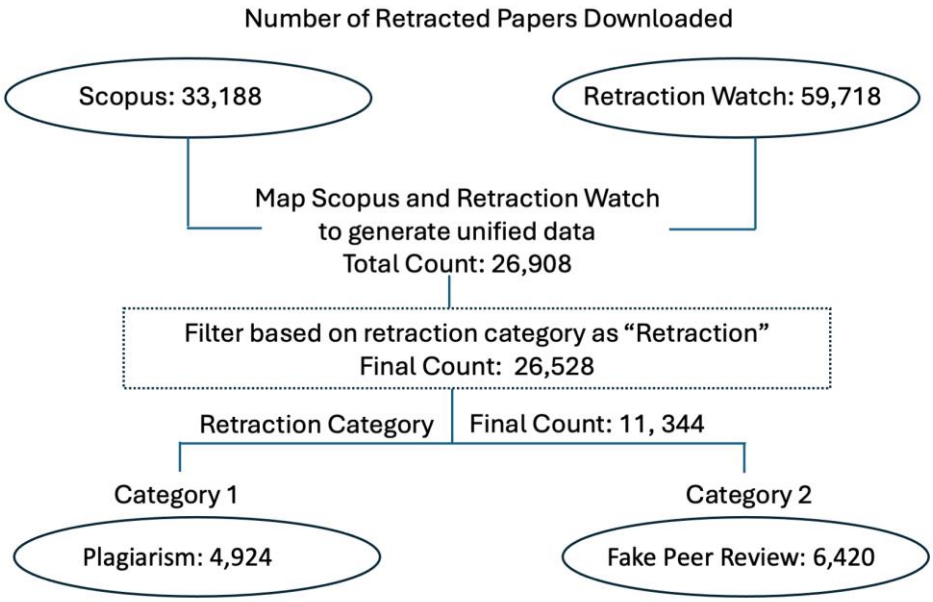


Figure 1. Data flowchart.

Retraction Categories

We classified the data into two categories: plagiarism and fake peer review, discarding the remaining data. These categories were chosen because of the clear documentation of retraction reasons and the significant number of retractions within them. The plagiarism category includes papers retracted for retraction reasons such as "Plagiarism of articles, data, images and texts + Duplication of article, data, image and text" and the category fake peer review includes papers retracted for retraction reasons such as "Fake peer review + Concerns / Issues with peer review" (https://retractionwatch.com/retraction-watch-database-user-guide/retraction-watch-database-user-guide-appendix-b-reasons/). Out of 26,528 filtered documents, we further categorized the data into two retraction categories such as Plagiarism -

category 1 and Fake peer review - category 2. We filter 4,924 classified as plagiarism and 6,420 as fake peer review. A total of 156 papers with 1,954 citations appeared with both retraction reasons; hence for simplicity, we excluded these papers. Table 1 shows the final count of retractions.

Table 1. Retraction count along with citations for both categories.

Retraction Category	Total Retractions	With Citations	Without Citations	Total Citations
Plagiarism	4,924	4,482	442	1,41,891
Fake Peer Review	6,420	5,197	1,223	55,272
Total	11,344	9,679	1,665	1,97,163

Results and Discussion

Publication and Retraction Trend

The number of retracted articles has increased significantly over recent years, driven by various factors (Sharma, 2024; Steen et al., 2013). Figure 2(a) represents the trend of publications and retractions over the years due to plagiarism. Publications show a steady rise from 2005, peaking in 2020, and then declining. The decline after 2020 might reflect stricter plagiarism detection. The overlap in publication and retraction trends suggests a robust but delayed system to identify plagiarism. Retractions increase from 2010, peak in 2021, and then decrease gradually, mirroring the publication trend. The retractions have shown a steady increase over the years, with an average growth rate of 1.2 times.

Fake peer-review is a growing issue in academia (Hadi, 2016). Figure 2(b) represents the trend of publications and actions over the years due to fake peer review cases. Publications start increasing significantly around 2011 and peak in 2021, suggesting a rise in published works associated with fake peer review. Retractions follow a similar trend, but lag slightly, with a dramatic increase from around 2017, peaking in 2022, and then sharply declining. In 2021, the retraction rate increased by 8.3 times compared to 2020. In 2022, it rose to 3.6 times the rate of 2021. By 2023, the increase was 1.9 times, indicating a decline in the growth rate of retractions in recent years.

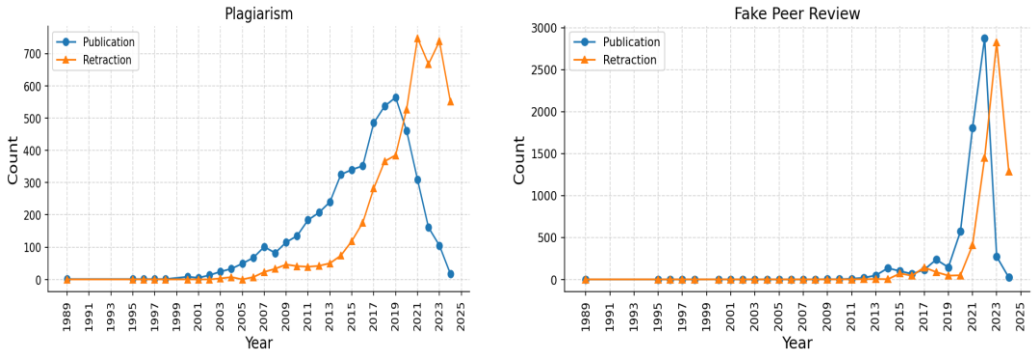


Figure 2. Number of papers published and later got retracted over years. (a) Plagiarism and (b) Fake peer review.

Figure 3 visualizes the retraction trends over a period of years for two categories: plagiarism and fake peer review. Plagiarism starts at a high percentage (around 23.2%) for papers retracted within the first year of publication. It gradually decreases over the years, showing that plagiarism cases are identified relatively early. However, fake peer review increases sharply in the first year, reaching over 42.4%, indicating that these cases are caught quickly after publication. The percentage drops steeply within the first few years and approaches zero after about 5 years, implying that this issue is typically resolved early. Overall, the fake peer review cases are identified and retracted more rapidly than plagiarism cases.

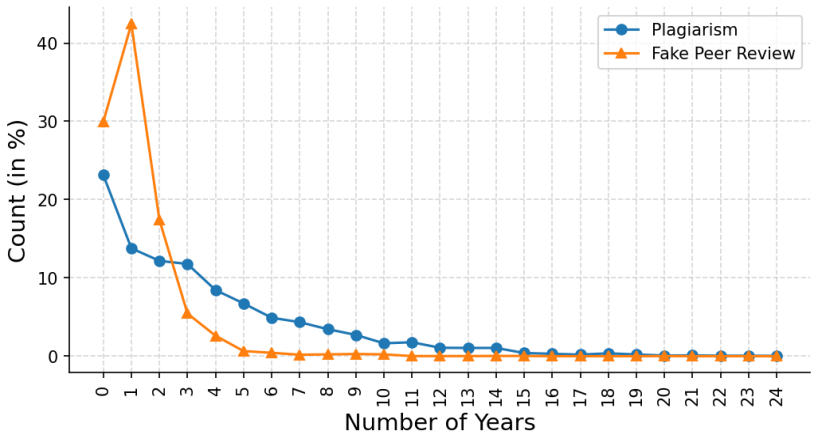


Figure 3. Retraction time.

Analysis of Retracted Citations

In 2009, the Committee on Publication Ethics (COPE) (Barbour et al., 2009) released retraction guidelines recommending that notices clearly explain the reasons for retraction and distinguish misconduct from honest error. These notices should be freely accessible and linked to the retracted article to prevent unintentional citations. A study of MEDLINE retracted articles (1966–1997) found that 94% of the citations

to retracted works were made unknowingly (Budd et al., 1998). Cassai et al. (2022) also highlighted in their study that 89% of the authors were unaware that they cited retracted articles which may be due to inadequate notification in journals and stored copies.

The results in Table 2 and Table 3 provide an overview of retracted and non-retracted citations across two retraction categories - Plagiarism and Fake Peer Review - and offer a detailed breakdown of retraction categories for retracted citations. Plagiarism accounts for a significantly higher total number of citations compared to fake peer review. Plagiarism has 98.4% non-retracted citations and 1.6% retracted citations, with 46.1% of its retracted citations attributed to plagiarism itself. Fake peer review has 97.6% non-retracted citations and 2.4% retracted citations, with 53.6% of its retracted citations attributed to fake peer review.

Table 2. Number of retracted and non-retracted citations in both categories.

Retraction Category	Total Citations	Number of Citations		Mapped with Retraction Watch
		NonRetracted	Retracted	
Plagiarism	1,41,891	1,39,621	2,270	2,100
Fake Peer Review	55,272	53,929	1,343	1,138
Total	1,97,163	1,93,550	3,613	3,238

Table 3. Retraction category of retracted citations.

Retraction Category	Retraction Category			Total Retracted Citations
	Plagiarism	Fake Peer Review	Others	
Plagiarism	967	76	1,057	2,100
Fake Peer Review	49	610	479	1,138
Total	1,016	686	1,536	3,238

Analysis of Self-citations

Self-citation in research refers to the practice in which authors cite their own previous works in new publications. Self-citations significantly contribute to the continued citation of retracted articles, and approximately 18% of the authors cite their own retracted work after retraction. There is also a positive correlation between self-citations and the total number of citations after retraction (Madlock-Brown and Eichmann, 2014). After analyzing Table 2 and Table 4, it is observed that only 1.49% of citations to retracted papers are self-citations in cases of plagiarism, and 1.96% in cases of fake peer review. Furthermore, when examining citations that were later retracted, 17.18% of these are self- citations in cases of plagiarism and 13% in cases of fake peer review. Overall, plagiarism involves fewer self-citations compared to fake peer review but has a higher proportion of retracted self-citations.

Table 4. Number of self-citations to retracted papers. A total of 3,205 self-citations are subset of 11,344 retracted papers.

Retraction Category	Self-citations		
	Total	Retracted	NonRetracted
Plagiarism	2,119	2,270	2,100
Fake Peer Review	1,086	1,343	1,138
Total	3,205	3,613	3,238

Furthermore, Table 5 presents the distribution of author pairs who self-cited their retracted papers. As shown in able 3, a total of 11,344 retracted papers received 197,163 citations. Of these, 8.5% were self-citations. Among the self-citations, individual authors (including repetitions) contributed 4.18%, with 70% of these citations directed toward papers categorized as plagiarized. Similarly, 2.13% of the citations were self-citations by pairs of authors (teams consisting of two authors, including repetitions but limited to groups of two). Within these 2.13% self-citations, most (76.79%) were citations to articles classified under the plagiarized category. In general, 75.4% of self-citations from different groups of authors were associated with plagiarized articles. This count of self-citations under plagiarized category keeps on increasing as the team size increases.

Table 5. Team size of authors who self-cited their retracted papers. A total of 16,871 self-citations is a subset of 197,163 citations received by 11,344 retracted papers.

Team Size	Self-citations				Team Size	Self-citations		
	Total	Plagiarism	Fake Peer Review			Total	Plagiarism	Fake Peer Review
1	8248	5778	2470		10	37	35	2
2	4201	3226	975		11	23	23	-
3	2084	1697	387		12	4	4	-
4	1080	914	166		13	4	4	-
5	608	516	92		14	6	5	1
6	262	237	25		15	4	4	-
7	167	152	15		17	1	1	-
8	91	81	10		19	1	1	-
9	50	47	3		**Total**	**16871**	**12725**	**4146**

Conclusion

Citing retracted studies is an important issue in academia because it risks spreading misinformation and undermining the integrity of the scientific literature (Van Noorden and Naddaf, 2024). Although retraction notices are issued, many retracted papers are still cited without noting their retracted status. This problem occurs in various fields, including computer science and biomedical research, where retracted papers are often cited in systematic reviews and meta-analyses. A major cause for this is that the authors are not informed about the status of article retractions, either

because they do not receive enough notifications in journals and databases or because they depend on saved copies or find uncorrected versions available on open-access platforms (Million and Budd, 2024).

The management and identification of retracted publications are challenging due to logistical issues and the decentralized nature of publication databases. Retraction notices are often not prominently displayed, and databases frequently fail to effectively link retracted articles to their corresponding notices. Improved visibility of retraction notices is essential, including clear labeling and alerts to prevent the continued citation of retracted articles. The authors must verify with diligence the status of the retraction of the articles they cite, and the reviewers must check that references in the manuscripts are current and correct.

The continued existence of citations to retracted articles, with a role played by self-citations, requires improved action practices and better awareness among scientists. An increased awareness of the consequences of citing retracted work, in conjunction with providing education on how to check the status of the article, can significantly reduce inappropriate citations (Cassai et al., 2023; Minetto et al., 2023). This will help ensure that the scientific literature remains credible and of high quality.

Data Availability Statement

The datasets used in the study will be available from the corresponding author on request.

Acknowledgment

The authors gratefully acknowledge the Research and Development Cell, BML Munjal University's financial support through the seed grant (No: BMU/RDC/SG/2024-06), which made this research possible.

Conflict of interest

The authors declare no conflict of interest.

References

Bar-Ilan, J., Halevi, G., 2017. Post retraction citations in context: a case study. Scientometrics 113, 547 – 565. doi:10.1007/s11192-017-2242-0.

Barbour, V., Kleinert, S., Wager, E., Yentis, S., 2009. Guidelines for retracting articles. committee on publication ethics; 2009 sep. doi:https://doi.org/10.24318/cope.2019.1.4.

Bell, K., Kingori, P., Mills, D.S., 2022. Scholarly publishing, boundary processes, and the problem of fake peer reviews. Science, technology & human values 49, 78 – 104. doi:10.1177/01622439221112463.

Bolboacă, S.D., Buhai, D., Aluaș, M., Bulboacă, A., 2019. Post retraction citations among manuscripts reporting a radiology-imaging diagnostic method. PLoS ONE 14. doi:10.1371/journal.pone.0217918.

Budd, J.M., Sievert, M., Schultz, T.R., 1998. Phenomena of retraction: reasons for retraction and citations to the publications. Jama 280, 296–297.

Cassai, A.D., Geraldini, F., Pinto, S.D., Carbonari, I., Cascella, M., Boscolo, A., Sella, N., Monteleone, F., Cavaliere, F., Munari, M., Garofalo, E., Navalesi, P., 2022.

Inappropriate citation of retracted articles in anesthesiology and intensive care medicine publications. Anesthesiology 137, 341 – 350. doi:10.1097/ALN.0000000000004302.

Cassai, A.D., Volpe, F., Geraldini, F., Dost, B., Boscolo, A., Navalesi, P., 2023. Citing retracted literature: a word of caution. Regional Anesthesia & Pain Medicine 48, 349 – 351. doi:10.1136/rapm-2022-104177.

Gray, R., Al-Ghareeb, A., Mckenna, L., 2019. Why articles continue to be cited after they have been retracted: An audit of retraction notices. International journal of nursing studies 90, 11–12. doi:10.1016/j.ijnurstu.2018.10.003.

Hadi, M.A., 2016. Fake peer-review in research publication: revisiting research purpose and academic integrity. International Journal of Pharmacy Practice 24. doi:10.1111/ijpp.12307.

Hagberg, J., 2020. The unfortunately long life of some retracted biomedical research publications. Journal of applied physiology doi:10.1152/ japplphysiol.00003.2020.

Hamilton, D.G., 2019. Continued citation of retracted radiation oncology literature—do we have a problem? International Journal of Radiation Oncology* Biology* Physics 103, 1036–1042.

Heibi, I., Peroni, S., 2021. A qualitative and quantitative analysis of open citations to retracted articles: the wakefield 1998 et al.'s case. Scientometrics 126, 8433–8470.

Heibi, I., Peroni, S., 2022. A protocol to gather, characterize and analyze incoming citations of retracted articles. Plos one 17, e0270872.

Kamali, N., abadi, A.T.B., Rahimi, F., 2020. Plagiarism, fake peer-review, and duplication: Predominant reasons underlying retractions of iran- affiliated scientific papers. Science and Engineering Ethics 26, 3455 – 3463. doi:10.1007/s11948-020-00274-6.

Khurana, P., Ganesan, G., Kumar, G., Sharma, K., 2022. A bibliometric analysis to unveil the impact of digital object identifiers (doi) on bibliometric indicators, in: Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021, Springer. pp. 859–869.

Koçyiğit, B., Akyol, A., Zhaksylyk, A., Seiil, B., Yessirkepov, M., 2023. Analysis of retracted publications in medical literature due to ethical violations. Journal of Korean Medical Science 38. doi:10.3346/jkms.2023.38.e324.

Kühberger, A., Streit, D., Scherndl, T., 2022. Self-correction in science: The effect of retraction on the frequency of citations. PLOS ONE 17. doi:10.1371/journal.pone.0277814.

Madlock-Brown, C., Eichmann, D., 2014. The (lack of) impact of retraction on citation networks. Science and Engineering Ethics 21, 127 – 137. doi:10.1007/s11948-014-9532-1.

Million, A.J., Budd, J.M., 2024. Disinformation in science: Ethical considerations for citing retracted works. Proceedings of the Association for Information Science and Technology doi:10.1002/pra2.1025.

Minetto, S., Pisaturo, D., Cermisoni, G., Rabellotti, E., Pagliardini, L., Candiani, M., Papaleo, E., Alteri, A., 2023. P-385 are you completely aware of your citations? a cross-sectional study on improper citations of retracted articles in medically assisted reproduction. Human Reproduction 38, dead093–349.

Palla, I.A., Singson, M., Thiyagarajan, S., 2023. Systematic examination of post- and pre-citation of indian-authored retracted papers. Learned Publishing 36. doi:10.1002/leap.1572.

Qi, X., Deng, H., Guo, X., 2016. Characteristics of retractions related to faked peer reviews: an overview. Postgraduate Medical Journal 93, 499 – 503. doi:10.1136/postgradmedj-2016-133969.

Rivera, H., 2018. Fake peer review and inappropriate authorship are real evils. Journal of Korean Medical Science 34. doi:10.3346/jkms.2019.34.e6.

Schneider, J., Ye, D., Ye, D., Hill, A.M., Whitehorn, A., Whitehorn, A., 2020. Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. Scientometrics 125, 2877 – 2913. doi:10.1007/s11192-020-03631-1.

Sharma, K., 2021. Team size and retracted citations reveal the patterns of retractions from 1981 to 2020. Scientometrics 126, 8363–8374.

Sharma, K., 2024. Over two decades of scientific misconduct in India: Retraction reasons and journal quality among inter-country and intra-country institutional collaboration. Scientometrics, 1–23.

da Silva, J.A.T., 2020. Reasons for citing retracted literature are not straightforward, and solutions are complex. Journal of applied physiology 129 1, 3. doi:10.1152/japplphysiol.00258.2020.

Silva, J.A.T., Bornemann-Cimenti, H., 2016. Why do some retracted papers continue to be cited? Scientometrics 110, 365–370. doi:10.1007/s11192-016-2178-9.

Steen, R., Casadevall, A., Fang, F., 2013. Why has the number of scientific retractions increased? PLoS ONE 8. doi:10.1371/journal.pone.0068397.

Tang, B., 2023. Some insights into the factors influencing continuous citation of retracted scientific papers. Publ. 11, 47. doi:10.3390/ publications11040047.

Van Noorden, R., Naddaf, M., 2024. Exclusive: the papers that most heavily cite retracted studies. Nature 633, 13–15.

Wang, D., Chen, S., 2025. An empirical study of retractions due to honest errors: Exploring the relationship between error types and author teams. Journal of Informetrics 19, 101600.

Wang, P., Su, J., 2022. Expert-recommended biomedical journal articles: Their retractions or corrections, and post-retraction citing. Journal of Information Science 50, 17 – 34. doi:10.1177/01655515221074329.

https://doi.org/10.51408/issi2025_065

Revisiting the Field Normalization Approaches/Practices

Xinyue Lu[1], Li Li[2], Zhesi Shen[3]

[1]*luxinyue@mail.las.ac.cn*
Chinese Academy of Sciences, National Science Library; University of Chinese Academy of Sciences, Department of Information Resources Management, 33 Beisihuan West Road, 100190 Beijing (China)

[2]*lili2020@mail.las.ac.cn*, [3]*shenzhs@mail.las.ac.cn*
Chinese Academy of Sciences, National Science Library, 33 Beisihuan West Road, 100190 Beijing (China)

Abstract

Field normalization plays a crucial role in scientometrics to ensure fair comparisons across different disciplines. In this paper, we revisit the effectiveness of several widely used field normalization methods. Our findings indicate that source-side normalization (as employed in SNIP) does not fully eliminate citation bias across different fields and the imbalanced paper growth rates across fields are a key factor for this phenomenon. To address the issue of skewness, logarithmic transformation has been applied. Recently, a combination of logarithmic transformation and mean-based normalization, expressed as ln(c+1)/mu, has gained popularity. However, our analysis shows that this approach does not yield satisfactory results. Instead, we find that combining logarithmic transformation (ln(c+1)) with z-score normalization provides a better alternative. Furthermore, our study suggests that the best performance is achieved when combining both source-side and target-side field normalization methods.

Introduction

Citation and its derivative indicators are commonly used to reflect impact and are among the most important quantitative metrics in scientific evaluation (Garfield, 2006). However, differences in citation potential among fields result in field biases in citation-based indicators (Leydesdorff & Bornmann, 2011). The development and improvement of metrics which support cross-field comparison become a crucial issue in scientometrics.

Citation field normalization encompasses two important problems: how to treat the field difference and how to conduct the normalization. As for the first problem, there are two main streams of research aimed at addressing field bias: source-side normalization and target-side normalization.

Theoretical basis of source-normalized methods is that the varying citation density across fields is due to differences in the length of references (Mingers & Yang, 2017; Zitt & Small, 2008). In 2008, Zitt and Small proposed to normalize the raw citation by considering the reference length of citing source（1/r） (Zitt & Small, 2008). Later the concept of active reference(1/a) is introduced in 2011 (Leydesdorff & Bornmann, 2011) and journal's activity factor in 2012 (Waltman et al., 2013; Waltman & van Eck, 2013b), to account for the different accumulation rates of citations across different fields. The prerequisites for source-normalized methods to function fully are overly idealized and cannot be achieved in practice. Waltman

(Waltman & van Eck, 2013a) conducted a systematic large-scale empirical comparison among three source-normalized methods, but the evaluation framework he used does not support statistical tests. Meanwhile, how topic growth relates to citation counts and impacts citing-side normalization (Leydesdorff & Opthof, 2010; Waltman et al., 2013; Waltman & van Eck, 2013b) is not intuitive and still not well understood (Sjögårde & Didegah, 2022). This gap highlights the need for further research to explore how topic growth dynamics influence citation patterns and normalization practices.

The primary idea behind target-normalized methods is to calculate a relative citation performance given a comparable set for each publication or journal, which is commonly based on a field classification system (Leydesdorff & Bornmann, 2011). Therefore, the first issue of this normalized approach lies in the selection of field classification system (Bornmann, 2020). Recent studies have shown that a paper-level classification system performs better than journal-level classification system in reducing the citation bias (Ruiz-Castillo & Waltman, 2015; Shu et al., 2019; Strotmann & Zhao, 2010).

Once the classification system is determined, the second issue is selecting the normalization approaches which typically receives relatively less attention but is crucial. Currently, mean-based normalization (c/mu) (Abramo et al., 2012a, 2012b; Radicchi et al., 2008) and z-score transformation (c-mu/std) are the widely used practices because they are intuitive and simple. Recently, the log transformation of citation is introduced to overcome the skewness of citation distribution (Brzezinski, 2015; Eom & Fortunato, 2011; Lundberg, 2007; Shen et al., 2018; Stringer et al., 2008). Furthermore, the normalization is applied to the transformed citation, especially the z-score normalization approach (Lundberg, 2007). A more detailed discussion of normalization approaches, including their applications and limitations, can be found in the review in 2016 (Waltman, 2016). Here leads to our second question, does this combination result better normalization performance (log-transformation + Z-score), or which type combination performs better?

In this paper, we want to answer the following questions:

(1) Can the source-normalized methods entirely eliminate citation bias among fields?

(2) For target-side normalization, among c/mu, c-mu/std, ln(c)/mu, ln(c)-mu/std, which approach has better performance?

(3) Will the combination of source-side and target-side normalization achieve better performance?

Data and Methods

Publication data and citation data

We collect articles and reviews indexed in the Web of Science (WoS) between 2020 and 2021 and their citations received in 2022. To ensure consistency in the data coverage, we focus exclusively on articles and reviews indexed in the Science Citation Index Expanded (SCIE) and Social Sciences Citation Index (SSCI)

categories. Finally, the dataset for 2020 and 2021 comprises a total of 450,810 papers, while the dataset for 2022 includes 2,221,501 papers. Additionally, the citation relationships in 2022 contain 118,294,005 citations, with 16,329,497 of these citations referencing core papers published in 2020 and 2021.

Classification systems

In this study, we leverage two distinct classification systems to categorize the collected publications, ensuring a more robust and unbiased approach to field normalization and evaluation. Specifically, we align the papers in our dataset with both the CWTS paper-level classification system, which provides hierarchical classification across three granularity levels—micro-level, meso-level, and macro-level topics (Waltman & van Eck, 2012)—and the SciSciNet subfield classification system, which is derived from the MAG (Microsoft Academic Graph) dataset and consists of 292 specific subfields (Lin et al., 2023). This dual-classification strategy addresses the potential issue of bias that may arise when using a single classification system for both normalization and evaluation, thus avoiding the "athlete and referee" situation, where the same classification system influences both the standardization and assessment processes.

Among the collected publications, 90.9% of publications can be matched to CWTS classification systems and 97.9% of publications can be matched to SciSciNet subfield classification system through DOI. For the unmatched papers, we generate embeddings based on title and abstract using SPECTER (Cohan et al., 2020) and apply the k-nearest neighbor algorithm(KNN) to find the most related classifications.

Citation Indicators

Building on the normalization approaches discussed earlier, the next step is to define the key bibliometric metrics that will be used in our analysis. These indicators are essential for evaluating the impact and performance of scientific publications, with citations being the most fundamental and widely-used measures.

In this section, we categorize the normalization methods into three distinct types: source-side metrics, target-side metrics, and dual-side metrics. Each category offers different approaches to adjust for field-specific biases.

Unnormalized metric

Citation count, c. The citation count refers to the citations received by paper i in a given year.

Source-side normalized metrics

① First source normalized citation count, $sc^{(1)}$. The $sc_i^{(1)}$ value of paper i is calculated as:

$$sc_i^{(1)} = \sum_{i=1}^{c_i} (\frac{1}{r_i}),$$

where r_i is the length of reference list in the paper from i^{th} citation. $sc^{(1)}$ would suppress citation bias among fields from source theoretically (Waltman et al., 2013).

② Second source normalized citation count, $sc^{(2)}$. The value $sc_i^{(2)}$ of paper *i* is calculated as:

$$sc_i^{(2)} = \sum_{i=1}^{c_i} (\frac{1}{a_i}),$$

where a_i is the number of active references in the paper from which i^{th} citation generates. Active reference is defined as papers in Web of Science, falling into the time window of analysis year (Waltman & van Eck, 2013b; Zitt & Small, 2008). For example, the active reference length for the 2-year time window of publications in 2022 refers to the number of references publishing between 2020 and 2021.

③ Third source normalized citation count, $sc^{(3)}$. The $sc_i^{(3)}$ value of paper *i* is calculated as:

$$sc_i^{(3)} = \sum_{i=1}^{c_i} (\frac{1}{a_i \times p_i}),$$

where the definition of a_i is the same as $sc^{(2)}$ and p_i is the proportion of publications which contains at least one active reference among all publications in journal of i^{th} citing publication (Waltman et al., 2013).

For the above four metrics, we also calculate their logarithmic form: $\ln(c_i + 1)$, $\ln(sc_i^{(1)} + 1)$, $\ln(sc_i^{(2)} + 1)$ and $\ln(sc_i^{(3)} + 1)$, and respectively defined them as $c^{\ln}$, $sc^{(1)\ln}$, $sc^{(2)\ln}$ and $sc^{(3)\ln}$.

Target-side normalized metrics

For target-side normalized metrics, we consider two normalize approaches: relative ratio and z-score.

① Relative ratio, $ratio^f$, we define it as

$$ratio_i^f = \frac{m_i}{\mu^f},$$

where m_i refers to metric value of paper *i* and μ^f is average metric value of papers which belongs to the same field with paper *i*.

② z-score, z^f.We define it as

$$z_i^f = \frac{m_i - \mu^f}{\sigma^f},$$

where μ_i^f is average metric value of papers which belongs to the same field with paper *i* and σ^f is the standard deviation of metric value in field *f*.

Dual-side normalized metrics

By combining source-side and target-side normalization approaches, we have the dual-side normalized metrics as shown in Table 1.

Table 1 shows the total indicators we investigated in this work. We combine Citation count, c and three source-side metrics with two different normalized approaches, resulting in 24 metrics (Table 1). The structure of Table 1 can represent the categories to which the normalization methods used for each metric belongs (non-normalized, source-normalized, target-normalized or both).

Table 1. The combination of citation-based metrics and normalized approaches.

-		*None*		*Source side normalization*					
		Original	*Log*	*Original*			*Log*		
None	-	c	$c^{\ln}$	$sc^{(1)}$	$sc^{(2)}$	$sc^{(3)}$	$sc^{(1)\ln}$	$sc^{(2)\ln}$	$sc^{(3)\ln}$
Target side Normalization	Ratio	$R(c)$	$R(c^{\ln})$	$R(sc^{(1)})$	$R(sc^{(2)})$	$R(sc^{(3)})$	$R(sc^{(1)\ln})$	$R(sc^{(2)\ln})$	$R(sc^{(3)\ln})$
	Z-score	$Z(c)$	$Z(c^{\ln})$	$Z(sc^{(1)})$	$Z(sc^{(2)})$	$Z(c^{\ln})$	$Z(sc^{(1)\ln})$	$Z(sc^{(2)\ln})$	$Z(sc^{(3)\ln})$

Evaluation Methodology

Evaluating bias among fields

We use two methods to assess whether the metrics correct bias among fields. The first qualitative method is based on a simple intuition: mean of the metric values in every meso-topic with field normalization effect should not have an obvious positive correlation with citation count that have not been normalized. So we will conduct scatter plots for each metric using field normalization methods against citation count to observe the relationship between them.

The second quantitative method is grounded in the following assumption: if the rankings derived from a given metric are not biased across scientific fields, then the proportion of publications from each field within the top z% of ranked publications should match the proportion of that field in the entire dataset (Dunaiski et al., 2019; Vaccario et al., 2017). In other words, publications from each field should be evenly distributed across every ranking interval. To quantitatively assess this deviation, we adopt the evaluation standard d_{M} proposed by Vaccario (Vaccario et al., 2017). Specifically, we compute the distributional inequality between the observed field

representation in the top z% and the expected distribution under field-neutral conditions. The greater this discrepancy, the poorer the effect of field normalization. For a given metric m, the expected number of papers from subfield i in the top $z\%$ under perfect field normalization is $\mu_i^{(m)} = (z/100) \cdot K_i$, where K_i is the total paper numbers in subfield i. The observed count $k_i^{(m)}$ represents the actual representation of subfield i in the top $z\%$. Then we can quantify the overall field bias using the Mahalanobis distance (d_{M}):

$$d_M^{(m)} = \sum_{i=1}^{F} \frac{(k_i^{(m)} - \mu_i^{(m)})^2}{\sigma_i^2} \cdot (1 - \frac{K_i}{N}),$$

where $\sigma_i^2 = \gamma \cdot K_i \cdot (N - K_i)$ is the expected variance and N is the total papers in the dataset. The finite-population correction factor $\gamma = \frac{n \cdot (N-n)}{N^2 \cdot (N-1)}$ accounts for the reduced variance in sampling without replacement, ensuring cross-sample comparability of bias measurements. The term $(1 - \frac{K_i}{N})$ dampens the disproportionate influence of dominant subfields on the aggregate bias metric, preventing overestimation from majority fields.

The 95% confidence interval for the simulated unbiased selection process using all publications represents the minimum standard to accomplish the task of field normalization, and d_{M} based on citation count represents a benchmark with no effect at all. It is worth noting that we utilized the micro-level topics from CWTS to standardize various metrics on the target side, while meso-level topics was employed to compute d_{M} to evaluate the effectiveness of the standardization. Additionally, we also used the subfield classification system from SciSciNet to recalculate d_{M} as a robustness check.

Benchmark of quantitative evaluation

We analyse the distribution of d_{M} using subfield classification of SciSciNet through a simulated unbiased selection process as a statistical null model. Specifically, we extract 10% of the total publications to calculate $\mathrm{d_M}$. Figure 1 illustrates the distribution of d_{M} with 500,000 simulations, with the upper bound of the 95% confidence interval estimated to be approximately 329.83. All rankings generated by the metrics described above will compute d_{M} and compare it with the value of 329.83.

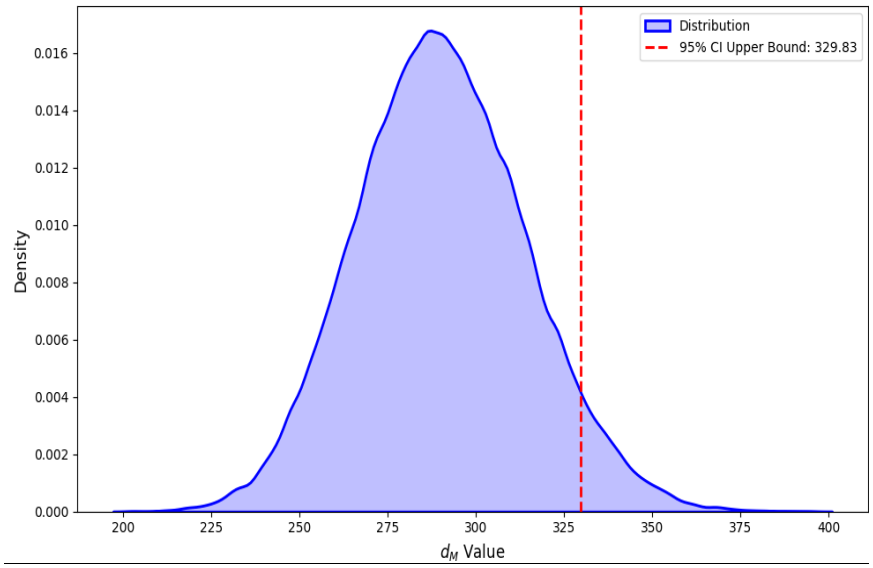


Figure 1. the distribution of d_{M}.

Results

RQ1: Can the source-normalized methods entirely eliminate citation bias among fields?

To address the question of which one of source-normalized methods can better correct the citation bias among fields, we construct scatter plots of several original metrics ($sc^{(1)}$, $sc^{(2)}$ and $sc^{(3)}$) against citation count c, without applying any normalized methods (Figure 2(a) – (c)). Among these, $sc^{(1)}$ with the smallest slope shows the best performance, but the differences in performance among the three source-normalized methods are not significant.

To better account for the influence of outliers and reflect the overall relationship between indicators, we rank the papers based on the values of the indicators and calculate the correlation among the rankings (Figure 2(d) – (f)). In ranking correlations, $sc^{(3)}$ exhibits the most effective correction for field bias. However, all three source-normalized methods ($sc^{(1)}$, $sc^{(2)}$ and $sc^{(3)}$) still show a strong positive correlation with citation count (c), suggesting that none of the three source-normalized indicators fully eliminate the field biases. Overall, $sc^{(3)}$ demonstrates the best performance in addressing citation bias among three source-normalized methods.

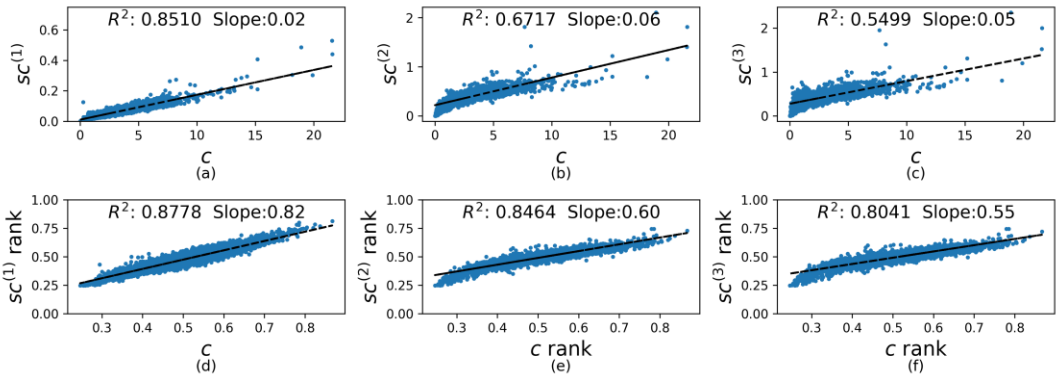


Figure 2. Correlation between citation count and source-normalized metrics.

We further validate the conclusion through a quantitative evaluation method based on d_M. The smaller the d_M, the better the normalization effect of metric among fields. As shown in Figure 3, $sc^{(3)}$ achieves the smallest d_{M} value across all percentiles, followed by $sc^{(2)}$, and then $sc^{(1)}$. All three methods perform better than the benchmark c, demonstrating a certain degree of effectiveness of the source-normalized methods in reducing field bias.

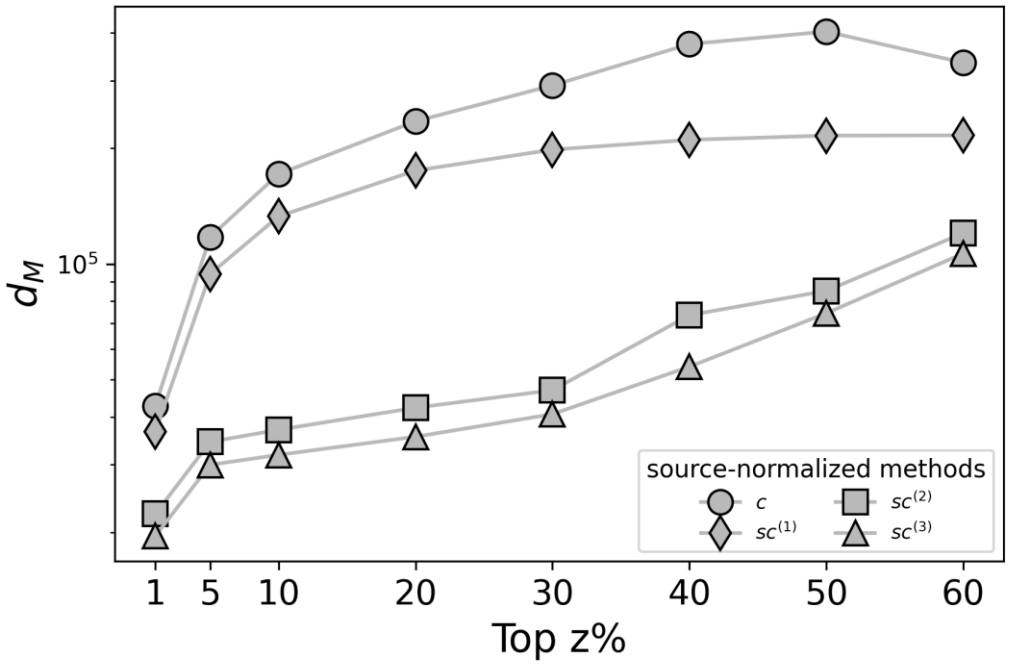


Figure 3. Field bias of source-normalized metrics.

We further investigate why source-side normalization methods fail to fully eliminate field bias in the normalization process. According to Waltman's 2013 paper on SNIP

(Waltman et al., 2013), there are three key assumptions for ensuring the effectiveness of source-side normalization:

(1) the same number of papers are published annually within each field, i.e., $M^f_{2020} = M^f_{2021} = M^f_{2022}$;

(2) there is no citation overlap between journals from different fields;

(3) each journal has at least one paper with an active reference.

If these three assumptions hold, the mean value (μ) of $sc^{(3)}$ for each field can be calculated as shown in the following formula (for details, see paper (Waltman et al., 2013)):

$$\mu = \frac{2(M_{2020} + M_{2021})}{M_{2022}} = 1.$$

The first two assumptions are difficult to achieve in practice and may help explain why these metrics fail to perform as expected. To further explore this issue, we test the validity of the first two assumptions.

The core of assumption 1 is that $M_{2022} = \frac{1}{2}(M_{2020} + M_{2020})$, implying that the number of papers published in a given field in 2022 should be equal to half of the total number of papers published in 2020 and 2021. However, in reality, the number of papers published in each field fluctuates every year, with varying degrees of change across different fields. This variation results in a mean value for $sc^{(3)}$, that deviates from 1. To quantify this variation, we define the *growth rate* of a field as $\frac{(M_{2020} + M_{2021})}{M_{2022}}$. A higher growth rate corresponds to a larger μ value for that field.

Assumption 2 posits that there is no citation overlap between journals from different fields. The core of source-normalized methods is the adjustment of citation counts by dividing them by the citation density of the corresponding field. If a paper is cited by journals from other fields, the citation density is either overestimated or underestimated, leading to normalization failure. To test this assumption, we define the *citation density* of a field.

The citation density of a field f, denoted as D_f, is defined as the total number of active references generated by all papers within the field. For a given paper i, its actual citation density AD_i is calculated as the weighted average of the citation densities of the fields that cite it:

$$AD_i = \sum_{k \in CitingFields_i} w_{i,k} \cdot D_k,$$

where $w_{i,k}$ is the proportion of citations received by paper i from field k and D_k is the citation density of field k. The expected citation density of paper i is the citation density of its own field f, denoted as D_f. Based on these, the density ratio for paper i is defined as:

$$DR_i = \frac{AD_i}{D_f},$$

this ratio greater than 1 indicates that the citation density of paper i is overestimated, potentially underestimating the paper's true impact. Conversely, a ratio less than 1 suggests that the citation density is underestimated, potentially overestimating the paper's impact.

Figure 4(a) shows the mean value of $sc^{(3)}$ for each meso-field in the CWTS classification against the growth rate. The colour of each data point represents the citation density ratio. We observe a clear positive correlation between the growth rate and the mean $sc^{(3)}$, and we find that when the citation density ratio is higher, the mean $sc^{(3)}$ tends to be smaller. Figure 4(b) demonstrates a positive correlation between the mean value of $sc^{(3)}$ and the growth rate, while showing a negative correlation with the citation density ratio. Further residual analysis reveals that the growth rate explains 63.7% of the variance in the mean value of $sc^{(3)}$, suggesting that it is a primary factor contributing to the failure of field normalization.

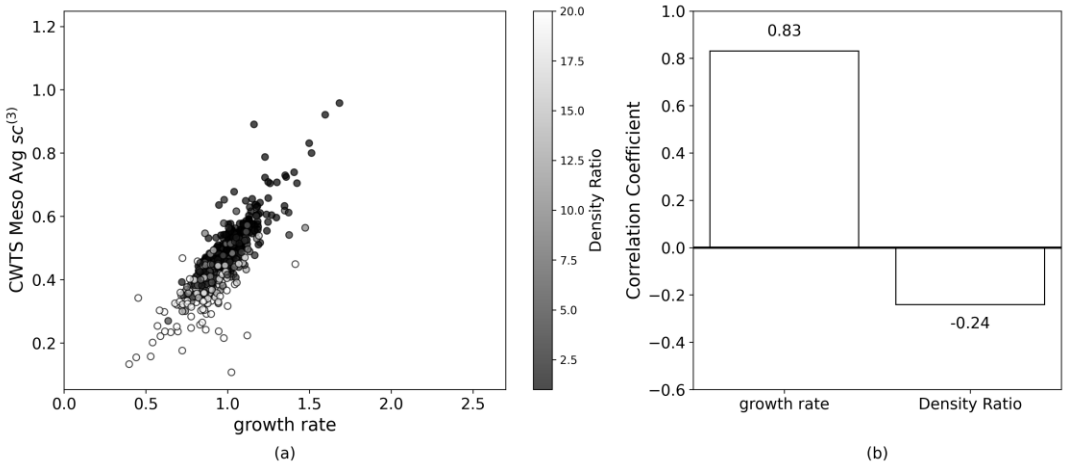


Figure 4. Factors affecting the effectiveness of source-side normalization. (a) Correlation between growth rate and average c for CWTS meso fields. (b) Strength of Correlation between Growth Rate/Density Ratio and Average $sc^{(3)}$ by CWTS meso fields.

RQ2: Among c/mu, c-mu/std, ln(c)/mu, ln(c)-mu/std, which approach has better performance?

To explore this question, we calculate the original form c, ratio-normalized original form $R(c)$, z-score-normalized original form $Z(c)$, ratio-normalized logarithmic form $R(c^{\ln})$, and z-score-normalized logarithmic form $Z(c^{\ln})$. According to the recommendation of previous research, we conduct the field normalization at the micro-level. Meanwhile, we evaluate the normalization performance at both CWTS meso-level and SciSciNet subfields. As shown in Fig.5, the results suggest that, under both evaluation schemes, no single method consistently outperforms others across all scenarios. However, overall, retaining the original citation counts and applying ratio normalization ($R(c)$) or using the logarithmic form combined with z-score normalization ($Z(c^{\ln})$) tend to yield relatively better outcomes.

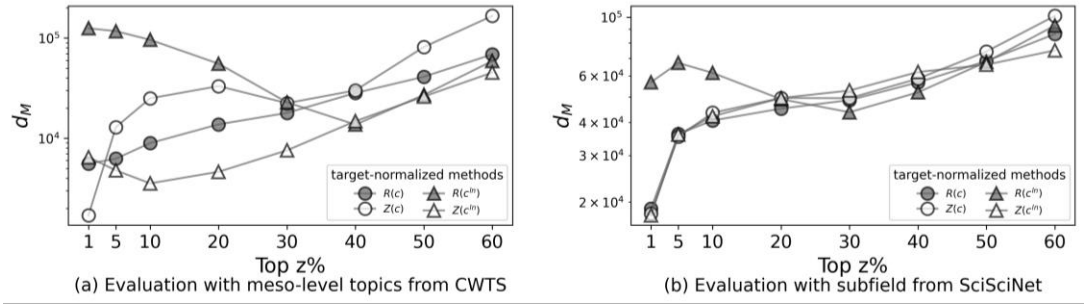


Figure 5. Field bias of different normalization approaches.

Strictly proving the effectiveness of these normalization metrics is challenging, but we can provide an intuitive explanation. Citation distributions are often approximated as log-normal distributions(Stringer et al., 2008). Under the logarithmic transformation, there is a natural connection between log(c) - μ and log(c/μ), leading to similar performance for $R(c)$ and $Z(c^{\ln})$. Additionally, since the variance across distributions is also considered for $Z(c^{\ln})$, the normalization performance is further improved. However, for log(c), which is already approximately normally distributed, using log(c)/μ, while aligning the means across different fields, tends to amplify the variance in fields with smaller means. This amplification gives these fields an advantage in top rankings and decreases the normalization performance.

RQ3: Will the combination of source-side and target-side normalization achieve better performance?

To address the question of whether combining source-side and target-side normalization can yield better performance than using source-side normalization alone, we leveraged the conclusions from *RQ1* and *RQ2*. *RQ1* demonstrated that $sc^{(3)}$ is the most effective source-normalized metric, while *RQ2* showed that applying ratio normalization to the original citation counts or using log-transformed z-score normalization generally yields better results. Building on these findings, we combined with the ratio normalization and log-transformed z-score methods to create two new indicators: $R(sc^{(3)})$ (ratio-normalized) and $Z(sc^{(3)\ln})$ (log-transformed z-score-normalized). These newly constructed indicators were then compared against existing indicators, including c, $R(c)$, and $Z(c^{\ln})$, to evaluate their relative effectiveness in normalizing citation data and reflecting a paper's impact within its field.

As illustrated in Figure 6(a), evaluation using the meso-level topics from CWTS indicates that combining source-side normalization with target-side normalization methods yields better results than using target-side normalization alone, with the Z-score method demonstrating superior performance for the field normalization task. Figure 6(b) presents the combination of source-normalized and target-normalized

methods ($R(sc^{(3)})$ and $Z(sc^{(3)\ln})$), demonstrating significantly better performance compared to other single-method approaches when subfields from SciSciNet was used as evaluation classification system. Among these, the combination of z-score normalization with the logarithmic form of $sc^{(3)}$, represented as $Z(sc^{(3)\ln})$, always emerges as the most effective.

These findings underscore the advantages of integrating source-side and target-side normalization methods. By leveraging their complementary strengths, the combined metrics provide a more effective and robust solution for addressing field-specific biases.

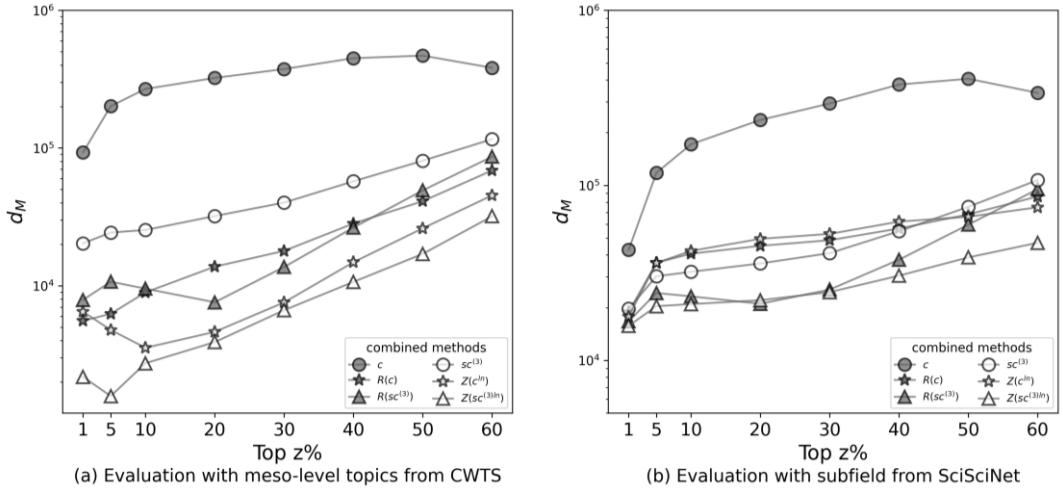


Figure 6. Field normalization performance for combing source and target-side approaches.

Conclusion

In this paper, we evaluated various source-normalized methods and found that while they achieve some success in reducing bias across fields, they are all unable to fully eliminate it. Our analysis, including residual analysis, indicates that imbalanced paper growth rates across fields are a key factor contributing to the limitations of these methods, which not only addresses the puzzle of why these methods are unable to fully eliminate field bias (Sjögårde & Didegah, 2022) but also opens avenues for future research to develop more refined normalization approaches that can better account for such dynamic factors.

We also found that using ratio normalization on original citation counts and log-transformation followed by z-score normalization both yields relatively strong results. However, directly applying ratio normalization after log-transformation is not a theoretically sound method. As a result, some studies that rely on this method should critically re-evaluate their findings.

Furthermore, by combining source-normalized and target-normalized methods, we found that the indicators constructed with ratio normalization ($R(sc^{(3)})$) and log-

transformed z-score normalization ($Z(sc^{(3)\ln})$) demonstrated relatively better performance compared to single-method approaches. However, these combinations still do not fully eliminate field differences within the 95% confidence interval. This suggests that while these combinations show promise, further refinement is needed to reduce biases more effectively.

These findings offer insights for the practical application of field normalization. Developing more robust normalization evaluation frameworks and exploring more effective ways to combine source-side and target-side normalization methods, along with their mathematical justification, will be crucial for enhancing the comparability across disciplines and improving citation metrics evaluation.

Acknowledgments

We thank Yahui Liu for valuable discussion.

References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012a). How important is choice of the scaling factor in standardizing citations? *Journal of Informetrics*, *6*(4), 645–654. https://doi.org/10.1016/j.joi.2012.07.002

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012b). Revisiting the scaling of citations for research assessment. *Journal of Informetrics*, *6*(4), 470–479. https://doi.org/10.1016/j.joi.2012.03.005

Bornmann, L. (2020). How can citation impact in bibliometrics be normalized? A new approach combining citing-side normalization and citation percentiles. *Quantitative Science Studies*, *1*(4), 1553–1569. https://doi.org/10.1162/qss_a_00089

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *SCIENTOMETRICS*, *103*(1), 213–228. https://doi.org/10.1007/s11192-014-1524-z

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D. S., & Assoc Computat Linguist. (2020). *SPECTER: Document-level Representation Learning using Citation-informed Transformers* (WOS:000570978202051). 2270–2282.

Dunaiski, M., Geldenhuys, J., & Visser, W. (2019). On the interplay between normalisation, bias, and performance of paper impact metrics. *JOURNAL OF INFORMETRICS*, *13*(1), 270–290. https://doi.org/10.1016/j.joi.2019.01.003

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and Modeling Citation Dynamics. *PLOS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0024926

Leydesdorff, L., & Bornmann, L. (2011). How Fractional Counting of Citations Affects the Impact Factor: Normalization in Terms of Differences in Citation Potentials Among Fields of Science. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, *62*(2), 217–229. https://doi.org/10.1002/asi.21450

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, *61*(11), 2365–2369. https://doi.org/10.1002/asi.21371

Lin, Z., Yin, Y., Liu, L., & Wang, D. (2023). SciSciNet: A large-scale open data lake for the science of science research. *SCIENTIFIC DATA*, *10*(1). https://doi.org/10.1038/s41597-023-02198-9

Lundberg, J. (2007). Lifting the crown—Citation z-score. *Journal of Informetrics*, *1*(2), 145–154. https://doi.org/10.1016/j.joi.2006.09.007

Mingers, J., & Yang, L. (2017). Evaluating journal quality: A review of journal citation indicators and ranking in business and management. *European Journal of Operational Research*, *257*(1), 323–337. https://doi.org/10.1016/j.ejor.2016.07.058

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272. https://doi.org/10.1073/pnas.0806977105

Shen, Z., Yang, L., & Wu, J. (2018). Lognormal distribution of citation counts is the reason for the relation between Impact Factors and Citation Success Index. *Journal of Informetrics*, *12*(1), 153–157. https://doi.org/10.1016/j.joi.2017.12.007

Sjögårde, P., & Didegah, F. (2022). The association between topic growth and citation impact of research publications. *Scientometrics*, *127*(4), 1903–1921. https://doi.org/10.1007/s11192-022-04293-x

Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2008). Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. *PLOS ONE*, *3*(2). https://doi.org/10.1371/journal.pone.0001683

Vaccario, G., Medo, M., Wider, N., & Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *JOURNAL OF INFORMETRICS*, *11*(3), 766–782. https://doi.org/10.1016/j.joi.2017.05.014

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Waltman, L., & van Eck, N. J. (2013a). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *JOURNAL OF INFORMETRICS*, *7*(4), 833–849. https://doi.org/10.1016/j.joi.2013.08.002

Waltman, L., & van Eck, N. J. (2013b). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *SCIENTOMETRICS*, *96*(3), 699–716. https://doi.org/10.1007/s11192-012-0913-4

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, *7*(2), 272–285. https://doi.org/10.1016/j.joi.2012.11.011

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, *59*(11), 1856–1860. https://doi.org/10.1002/asi.20880

https://doi.org/10.51408/issi2025_004

Role of Artificial Intelligence in Scientific Research: Classification Framework and Empirical Insights

Zhe Cao[1], Yuanyuan Shang[2], Lin Zhang[3], Ying Huang[4], Gunnar Sivertsen[5]

[1] *caozhe@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)

[2] *shangafra@163.com*
Chinese Academy of Social Sciences Evaluation Studies, Beijing (China)

[3] *linzhang1117@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)

[4] *ying.huang@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)

[5] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo (Norway)

Abstract

Artificial intelligence (AI) represents the culmination of multidisciplinary scientific knowledge and, through its diverse technological capabilities, has significantly contributed to research across various fields. This study explores the bidirectional relationship between AI and scientific research, with a focus on the frequency and context in which AI is mentioned in research articles. A classification framework is developed to categorize the different ways AI is mentioned in articles. Empirical analysis is conducted in the fields of oncology, nanoscience and nanotechnology, as well as meteorology and atmospheric sciences. The findings indicate that while the mention of AI in research articles has become widespread, the distribution of different ways of mentions is relatively concentrated. We not only identify the conceptual differences in the focus of AI mentions, but also uncover the disparities in the intensity and ways of AI mentions across research articles from various countries. This study contributes to a deeper understanding of the complex interplay between AI and scientific research, advocating for the synergetic development of AI and scientific progress by leveraging the strengths of different nations.

Introduction

The 2024 Nobel Prizes have marked a significant milestone in the convergence of artificial intelligence (AI) and scientific research. This year, the Nobel Prize in Physics was awarded to two pioneers in the field of AI, John Hopfield and Geoffrey Hinton, who utilized concepts and methods from physics to invent the Hopfield

network and Boltzmann machine, respectively, laying the foundation for machine learning and artificial neural networks. Simultaneously, the Nobel Prize in Chemistry was granted to David Baker, Demis Hassabis and John Jumper for their revolutionary advancements in protein structure prediction, achieved through the innovative integration of AI with computational chemistry. These awards not only emphasize the multidisciplinary collaboration driving AI innovation but also highlight AI's pivotal role in enabling scientific breakthroughs. Together, they exemplify the bidirectional synergy between AI and science, advancing both to remarkable new frontiers.

This development raises an important question: What role does AI play in science? For a long time, researchers have recognized how AI integrates interdisciplinary scientific knowledge, emphasizing the contributions of various scientific fields to the development of AI (Arencibia-Jorge et al., 2022). In recent years, increasing attention has been given to the enabling role of AI in scientific progress, with its growing use and benefits across diverse scientific domains (Gao et al., 2024). Researchers have recognized the emergence of the 5th Paradigm of Science – AI-driven science (Ioannidis, 2024), and discussed various ways in which AI supports scientific research in different research fields (He, 2024; Wang et al., 2023). Still remaining, however, are systematic and quantitative empirical investigations into the bidirectional relationship between AI and science, particularly regarding how AI both shapes and is shaped by scientific advancements (Miao et al., 2022; Xu et al., 2024). Further exploration is needed of the various manifestations of this reciprocal relationship across research domains and among research entities. Such analyses may provide practical insights for enhancing the interaction between AI and science in real-world research, thereby promoting their parallel development.

To advance knowledge in this direction, this study addresses two central questions: (1) How can we understand and distinguish the various roles of AI in science? (2) Do these roles vary across temporal periods, scientific fields and actors? To answer these questions, we will focus on textual mentions of AI in research articles, thereby taking one of the primary outputs of scientific research as an entry point. Such mentions may provide evidence of the roles of AI as currently recognized within the academic communities. By developing a classification framework and conducting an empirical analysis, we aim to map the landscape of AI influences across scientific fields and derive meaningful insights from the findings.

Classification framework

Acknowledging the bidirectional relationship between AI and science, we classify research articles that mention AI into two primary categories: *Science for AI* (Type A) and *AI for Science* (Type B). The first category emphasizes the cross-disciplinary exchange and integration of scientific knowledge (Frank, 1988) and highlights the role of various scientific fields in supporting AI development, primarily reflecting the role of AI as the research subject. The second category draws on the theory of parallel intelligence (Miao et al., 2024) and suggests that AI can augment human capabilities in conducting scientific research in specific ways, mainly reflecting the

role of AI as the research tool. In addition to the two primary categories, we posit that there are other circumstances in which articles mention AI in a relatively inconsequential manner. Consequently, we establish a third category – *Other* (Type C), to encapsulate the more ambiguous relationships between AI and science. Below, we will elaborate on the subdivisions of the three categories and their meanings, as detailed in Table 1.

As for Type A, we distinguish different subtypes of *Science for AI* by examining the core issues encountered by AI throughout its various stages of development. As a general-purpose technology (Bresnahan et al., 1995), AI has developed a wide range of technological capabilities and permeated diverse industries and fields. Scientists are not only focused on the development and improvement of AI technology (e.g. establishing a deep learning model), but also on promoting its application in real-world scenarios (e.g. evaluating the effectiveness of using AI in disease diagnosis). Furthermore, there is an increasing necessity to recognize and address the unintended consequences that may arise from its deployment (e.g. investigating the issue of algorithmic bias). Following this understanding, we have identified three subtypes – *Science for the Development of AI* (Type A1), *Science for the Application of AI* (Type A2) and *Science for the Governance of AI* (Type A3).

As for Type B, we distinguish different subtypes of *AI for Science* by examining the various tasks that AI performs in scientific research. As AI is increasingly utilized in scientific research, several attempts have been made to characterize the ways in which AI enhances scientific pursuits. For example, the Royal Society (2024) has outlined three primary functions of AI in scientific research – a computational microscope, a resource for human inspiration, and an agent of understanding. The European Commission (2023) has summarized the most common applications of AI in the research process, including prediction problems, transformations of input data, optimal parameterization, literature review, literature-based discovery, and automation of tedious, routine laboratory tasks. A report released by Google DeepMind (2024) has pinpointed five opportunities to accelerate science with AI, namely knowledge, data, experiments, models and solutions. Wang et al. (2023) have reviewed the role of AI in scientific research from four aspects – AI-aided data collection and curation, learning meaningful representations of scientific data, AI-based generation of scientific hypotheses, and AI-driven experimentation and simulation. Messeri et al. (2024) have proposed four uses of AI in the research process – the use of AI as Oracle for the study design, as Surrogate for the data collection, as Quant for the data analysis, and as Arbiter for the peer review. By combining insights from such proposals for categorizations with our observations from empirical data (see Section 3), we have identified four subtypes – *AI for Data Collection* (Type B1), *AI for Data Representation* (Type B2), *AI for Generation* (Type B3), and *AI for Simulation* (Type B4).

As for Type C, we do not further subdivide the category. Research articles under this type may mention AI in specific contexts, but AI is not an indispensable component in the conduct of the research, thereby neither being the subject of study as in Type A nor serving as the research tool as in Type B. Nevertheless, this type of research

remains of notable importance. AI may be mentioned to provide a contextual backdrop, reflecting its status as a current hot topic, or it could be indirectly supported by the research, indicating a special and potential relationship between the research and AI. The characteristics of studies under this type will be further explored in the empirical analysis to reveal more nuanced insights.

Table 1. Classification framework of AI mentions in research articles.

Categories		*Description*
A. Science for AI	A1. Science for the Development of AI	This type of study provides theoretical or methodological support for the technological development, improvement, and application of AI.
	A2. Science for the Application of AI	This type of study provides an overview, comment or evaluation of the progress, dilemmas, challenges, and potentials in applying AI to solving the problems within certain fields.
	A3. Science for the Governance of AI	This type of study provides a discussion on ethical, legal and policy problems arising from AI technologies and possible solutions to these problems.
B. AI for Science	B1. AI for Data Collection	The application of AI technologies for gathering data for further processing and in-depth analysis, to solve the problems in certain research fields.
	B2. AI for Data Representation	The application of AI technologies for structuring, modeling, and feature extraction from data, to solve the problems in certain research fields.
	B3. AI for Generation	The application of AI technologies for emulating human reasoning and cognition to generate creative content by calculating and mining large datasets, to solve the problems in certain research fields.
	B4. AI for Simulation	The application of AI technologies for simulating experimental or real-world scenarios to conduct predictive analysis of potential situations and outcomes, to solve the problems in certain research fields.
C. Other		Studies where AI is mentioned without having an indispensable role in the research.

Empirical data

Sample selection

This study uses research articles mentioning AI in representative fields within years from 2014 to 2023 as samples[1]. The data source is Web of Science (WoS) Core Collection. We selected three fields representing different areas of research by using three WoS categories – *Oncology* (ON), *Nanoscience and Nanotechnology* (NN), and *Meteorology and Atmospheric Sciences* (MA) – for exploratory analysis.

In selecting the case fields, we have considered three aspects. Firstly, given the varying sizes of fields within the WoS categories, we need to focus on fields that are relatively targeted and of moderate granularity. The fields of ON, NN and MA meet our need, with the detailed descriptions of the samples provided in the following subsection. Secondly, these three fields cover natural sciences, engineering sciences, and life sciences, encompassing both basic and applied value, and are capable of reflecting the research characteristics of different disciplinary domains. Thirdly, a more important consideration is that these fields feature typical examples of the intersection between AI and scientific research, such as the AI-assisted cancer screening in the ON field (McKinney et al., 2020), the AI-driven material discovery in the NN field (Szymanski et al., 2023), and the AI-supported weather forecasting in the MA field (Bi et al., 2023). In November 2024, Google announced nine ways in which AI is advancing science. Several of them involve the three representative fields mentioned above, including protein structure prediction, accelerating materials science, saving lives with accurate flood forecasting, predicting weather faster and with more accuracy, etc. Promoting the application of AI in health, environment, climate, and other fields, has become a focal point for academia, industry and policymaking circles (CB Insights, 2024; OECD, 2023). Based on our bidirectional classification framework, we will further explore and reveal the complex interactive relationships between AI and scientific research in these three fields.

It should be noted that by focusing our study on research articles that mention AI, we may overlook some specific connections between AI and science. For instance, prior to the popularization of the term of AI, numerous research fields have laid foundational theoretical and methodological groundwork for the development of AI, such as probability and information theory in mathematics, genetic phenomena and neural networks in biology, etc. These efforts, in a broader sense, constitute a form of *Science for AI*, yet they fall outside the scope of this study. In addition, AI may assist researchers in tasks such as literature review, hypothesis formulation and academic writing. However, these contributions of AI are often not explicitly discussed in papers and, therefore, are not included in our analysis. Recognizing the

[1] On the one hand, we aim to obtain the latest data that reflects contemporary trends. Given that our research was conducted in 2024, the most recent year covered is 2023. On the other hand, we strive to encompass key milestones in the development of AI over recent years, such as the emergence of BERT in 2018, hence we have selected a time window spanning a decade.

boundaries of our research sample is essential for a proper understanding of the subsequent analyses and results of this research.

Search strategy

We use two approaches to retrieve articles mentioning AI. One is to select articles that explicitly mention the term "artificial intelligence" or its abbreviations in certain contexts (considering that the abbreviation "AI" can refer to different concepts across various disciplines) in the title, keywords, or abstract. These three information fields are the ones that WoS can provide directly reflecting the content of the articles. The other approach is to include terms referring to representative sub-technologies of artificial intelligence as search keywords. Considering the broad definition of AI, traditional machine learning techniques such as Naive Bayes are not the focus of this study. According to a survey targeting scientists across various academic disciplines worldwide, ChatGPT and its LLM cousins are the tools that researchers mentioned most often when asked to type in the most impressive or useful example of AI tools in science (van Noorden et al., 2023). To a certain extent, ChatGPT and LLM represent a landmark moment in the development of AI, with the potential to disrupt existing paradigms and, optimistically, exert a positive influence on humanity (Vert, 2023). Therefore, terms related to ChatGPT and large language models (LLM) are selected for inclusion in the search terms.

Based on the above two approaches, we have developed a set of search terms by referencing existing search queries (Arencibia-Jorge et al., 2022; Mariani et al., 2024), performing manual filtering and supplementary additions, as well as consulting experts in the relevant fields, as shown in Table 2. We conducted searches in the *Title* (TI), *Author Keywords* (AK) and *Abstract* (AB) fields of the WoS database. After excluding six papers with false positives by manual checking, a total of 1,251, 1,189 and 364 articles mentioning AI were retrieved in the fields of *Oncology*, *Nanoscience and Nanotechnology*, and *Meteorology and Atmospheric Sciences*, respectively. These articles were subsequently utilized as the analytical samples for this study. The data retrieval date is January 3rd, 2025.

Table 2. Search terms of research articles mentioning AI.

Search strategies	*Search terms*
Full name	"artificial intelligen*"
Abbreviations in certain contexts	"strong AI" OR "full AI" OR "human-level AI" OR "AI for science" OR "AI4S" OR "AI4Science" OR "generative AI" OR "AI-generated content*" OR "AIGC" OR "AI-based research"
Keywords for representative AI technologies	"large language model*" OR "generative language model*" OR "generative pretrained transformer*" OR "generative pretrained language model*" OR "ChatGPT*" OR "GPT-1*" OR "GPT-2*" OR "GPT-3*" OR "GPT-4*" OR "GPT-5*" OR "GenAI" OR "OpenAI GPT" OR "Midjourney*"

Data annotation

Based on the classification framework shown in Table 1, this study categorizes 2,794 research articles with abstracts through manual annotation, carried out by two authors with significant experience in data annotation. After thoroughly understanding the connotations and characteristics of different types of articles among three selected fields, the annotators read the abstracts independently and labeled each article with the type of mentioning AI. During the annotation process, the annotators prioritized sentences that mention AI or its sub-technologies. The type of article is determined by analyzing the key terms within these sentences and the surrounding context. If a single sentence mentioning AI or its sub-technologies does not provide enough information to classify the article, the classification is further refined based on the overall content of the abstract.

For the results of manual annotation, the Kappa consistency test was conducted using the SPSS software. The annotation results of two annotators in the three research fields are statistically consistent, as shown in Table 3, indicating that the annotation results are usable for further analysis. In cases where discrepancies in the annotation results arose, the two annotators discussed the articles together until a consensus was reached.

Table 3. Kappa statistics of manual annotation results.

Research field	*Kappa value*
Oncology	0.982***
Nanoscience and Nanotechnology	0.997***
Meteorology and Atmospheric Sciences	0.908***

Note: *** indicates that the result is statistically significant.

Results

Overview of articles mentioning AI

Figure 1 illustrates the number of articles mentioning AI and their proportion within the total number of articles in the fields of *Oncology* (ON), *Nanoscience and Nanotechnology* (NN), and *Meteorology and Atmospheric Sciences* (MA) from 2014 to 2023. It can be observed that, over time, the frequency of AI mentions in research articles has steadily increased, especially since 2018, reflecting the growing integration of AI in scientific research. Specifically, the proportion of articles mentioning AI in the ON field has shown a more significant increasing trend.

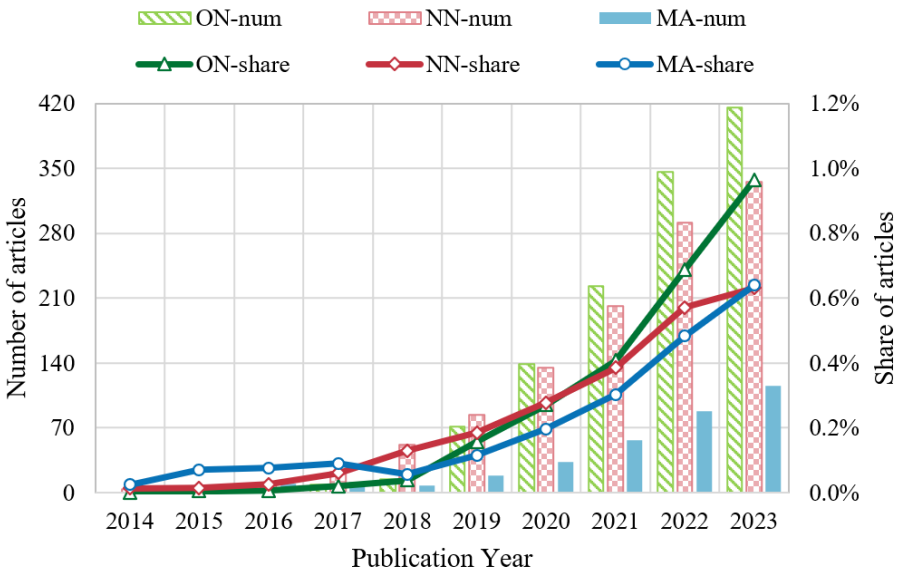


Figure 1. Number of articles mentioning AI and their proportion within the total articles in the corresponding fields.

Note: The full names and abbreviations of three fields are as follows – Oncology (ON), Nanoscience and Nanotechnology (NN), Meteorology and Atmospheric Sciences (MA).

When examining the different types of AI mentions, both commonalities and differences emerge across the three research fields, as shown in Figure 2. Our first observation is that the data representation is the primary way through which AI contributes to scientific research, with Type B2 articles accounting for 70%, 26% and 59% of the overall samples in the fields of ON, NN and MA respectively. At the same time, we observe that the ways AI is mentioned vary in emphasis across the three fields. Specifically, the ON field exhibits a relatively higher proportion of Type A2 research (15%), the NN field generates a substantial amount of Type C research (64%), and the MA field produces relatively more Type B4 research (18%).

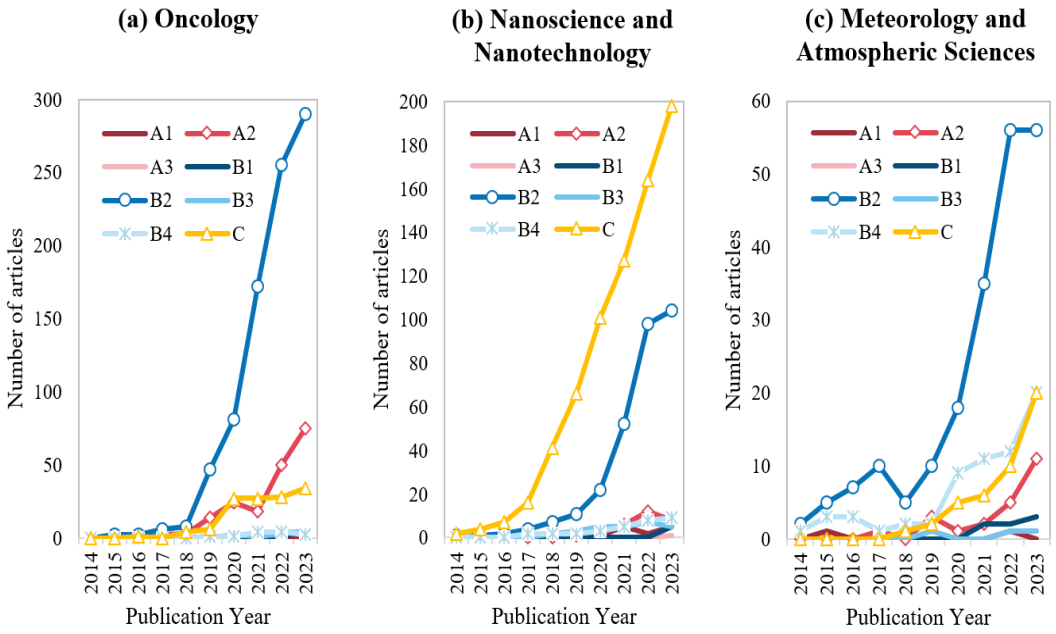


Figure 2. Number of articles with different AI-mention types in three research fields.

Thematic features of AI mentions

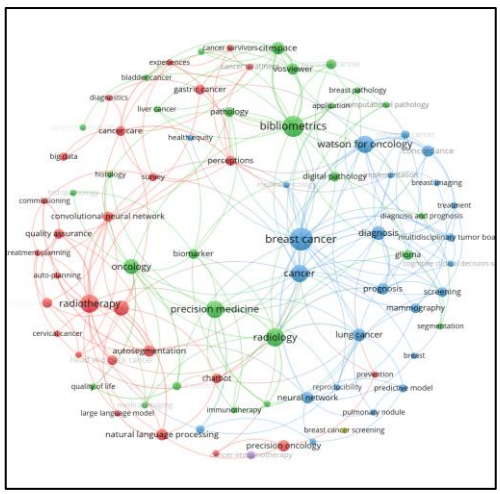
As AI is increasingly mentioned across a wide range of research articles, often by diverse ways, it is essential to investigate the specific problems these studies seek to address with the mention of AI. This section will analyze articles with representative types from three distinct fields, constructing keyword co-occurrence networks to identify and elucidate the core thematic features of these studies, as shown in Figure 3.

In the field of *Oncology* (ON), the primary focus is on Types A2 and B2. Type B2 constitutes the most prevalent category of articles within the field, whereas Type A2, though less dominant, exhibits a comparatively higher volume of publications relative to the other two fields. Type A2 (Science for the Application of AI) includes topics such as the bibliometric analysis on the research progress of AI usage in specific scenarios (green cluster), evaluating the effectiveness of AI applications (blue cluster), and investigating the attitudes of different stakeholders towards the use of AI (red cluster). These discussions aim to promote the more effective utilization of AI in real-world scenarios by examining the current status and impacts of AI usage. The prevalence of studies under this type reflects a cautious academic stance toward the application of AI in areas involving human health and life. Type B2 (AI for Data Representation) includes topics such as disease identification and classification (red cluster), organ segmentation (blue cluster), and tumor metastasis prediction (green cluster), primarily focused on medical image analysis and processing. The concentration on this type highlights that the ways of applying AI in oncology research are still relatively narrow in scope.

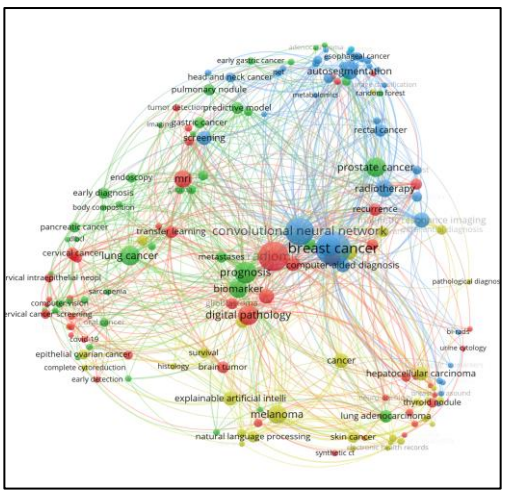
In the field of *Nanoscience and Nanotechnology* (NN), the primary focus is on Types B2 and C. These two types are the most prevalent categories of articles in this field, with the number of articles with the Type C being particularly prominent. Type B2 (AI for Data Representation) includes topics such as motion, odor and status detection (blue cluster), material structure analysis (green cluster) and related areas, though the topics in this category tend to be fragmented. In contrast, Type C (AI without a clear role in the research) encompasses topics such as artificial synapses (yellow cluster), memristors (blue cluster) and flexible materials (red cluster), with these technologies playing a significant role in advancing AI. For example, research on flexible materials can enhance the application of AI in sensor devices. Keyword analysis of a substantial body of Type C research reveals that the NN field is underpinned by robust foundational support, positioning it as a fundamental area within basic research for the development of AI. The inclusion of AI in studies within this field highlights the bidirectional relationship between AI and interdisciplinary knowledge, demonstrating how AI not only integrates insights from various disciplines but also drives the advancement of multiple fields.

In the field of *Meteorology and Atmospheric Sciences* (MA), the primary focus is on Types B2 and B4. While Type B2 (AI for Data Representation) represents the most prevalent category of articles within the field, Type B4 (AI for Simulation) exhibits a comparatively higher volume of publications relative to the other two fields. The topics arising from these two types of research are relatively consistent, both addressing issues such as weather forecasting (blue cluster in subgraph (e) and green cluster in subgraph (f)), disaster management (red clusters in subgraphs (e) and (f)), and air quality monitoring (green cluster in subgraph (e) and blue cluster in subgraph (f)). However, given the specialized nature of these topics, AI can contribute to addressing specific challenges through a range of methodologies, including data representation and simulation techniques. This underscores the diverse and multifaceted enabling capabilities of AI in advancing research within the MA field.

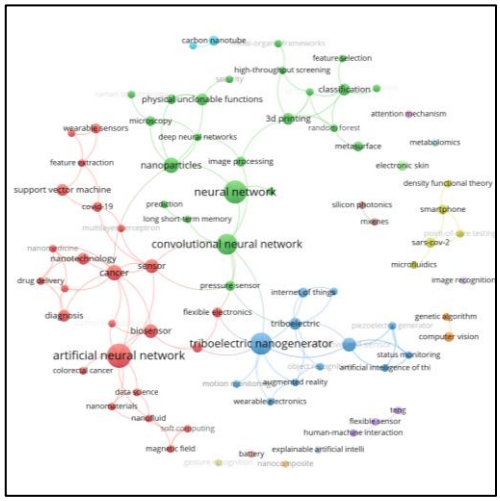
(a) ON-A2



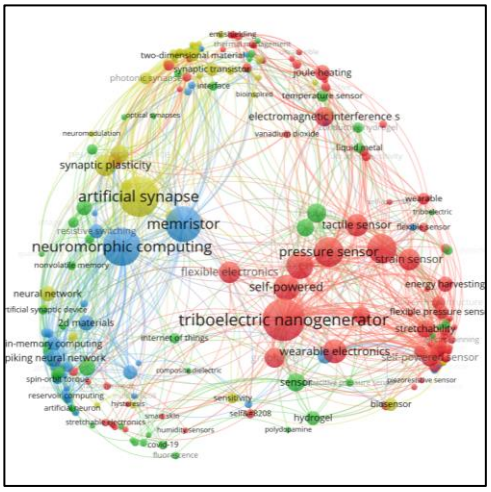
(b) ON-B2



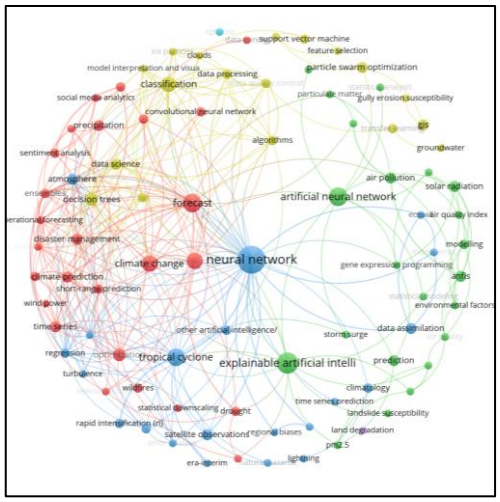
(c) NN-B2



(d) NN-C



(e) MA-B2



(f) MA-B4

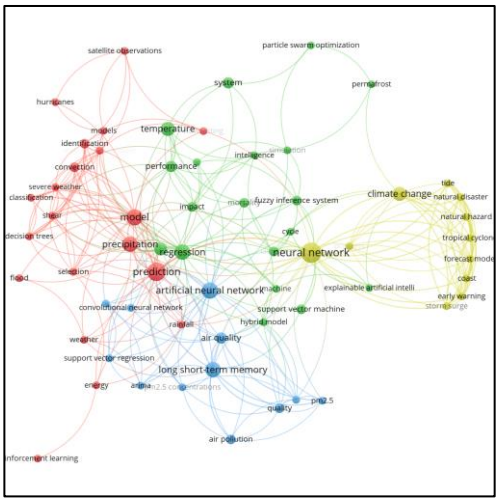


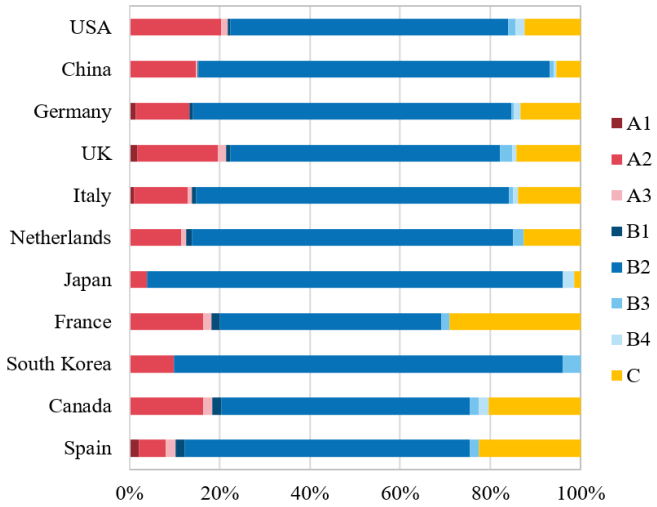
Figure 3. Keyword co-occurrence networks of articles with different types in three research fields.

Note: (1) The network is constructed using the VOSViewer software. (2) The nodes represent keywords; the size of each node indicates the frequency of keyword occurrence; nodes with different colors belong to different topic clusters; the thickness of the edges between nodes reflects the frequency of keyword co-occurrence. (3) Due to the limited number of articles with Type B4 in the MA field, Subgraph (f) uses both author keywords and WoS supplementary keywords, while the other subgraphs rely solely on author keywords. (4) The terms “artificial intelligence”, “machine learning” and “deep learning”, which are frequently occurring and highly generalizable technical keywords, are excluded to prevent overshadowing other thematic terms in the network.

Actor characteristics of AI mentions

Up to this point, we have observed that AI has been mentioned in research articles focused on various topics in different ways. The next question that arises is: who is conducting these studies? This section will examine the actor characteristics at the national level, with the aim of exploring the differences and similarities in the extent and specific ways that AI is mentioned in research articles across different countries. Figure 4 presents the distribution of the types of articles mentioning AI that were published in the field of *Oncology* (ON) by the eleven largest contributing countries[2], along with the proportion of articles mentioning AI within the total of articles from each country in the corresponding field. Our first observation is that, compared to countries such as the United States, the United Kingdom and Canada, which place greater emphasis on conducting research for advancing AI, countries like China, Japan and South Korea focus more on the specific utilization of AI in scientific research, with a more concentrated distribution of AI mentions within the Type B2. Our second observation is that, although China and the United States have the highest number of articles mentioning AI, European countries such as Germany, the United Kingdom, Italy, the Netherlands and Spain have higher proportions of articles mentioning AI in their total oncology articles.

(a) Type distribution of articles



(b) Statistics information

Country	Number	Share
USA	362	0.28%
China	351	0.24%
Germany*	150	0.51%
UK*	112	0.42%
Italy*	101	0.40%
Netherlands*	87	0.58%
Japan	77	0.23%
France	55	0.26%
South Korea	51	0.29%
Canada	49	0.26%
Spain*	49	0.36%

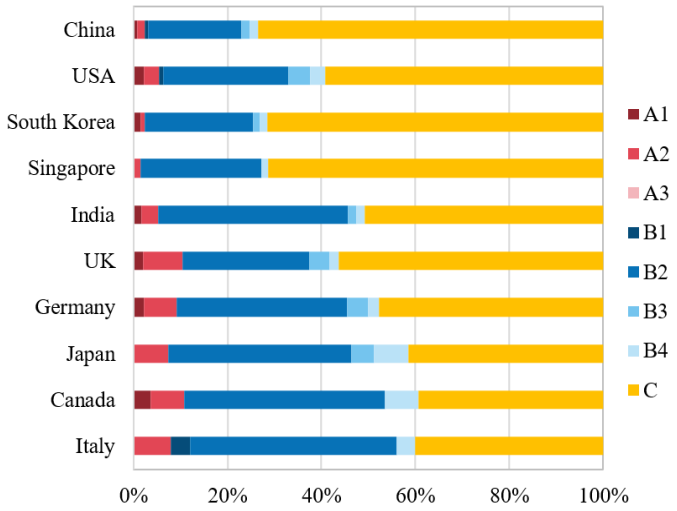
Figure 4. Number, distribution of types and share in all articles in the ON field of the largest contributing countries.

Note: * indicates that the proportion of articles mentioning AI in the total publications of the country is relatively prominent among all countries.

[2] Canada and Spain are tied for tenth in the number of papers, so both have been included.

Figure 5 presents the distribution of the types of articles mentioning AI published in the field of *Nanoscience and Nanotechnology* (NN) by the ten largest contributing countries, along with the proportion of articles mentioning AI within the total of articles from each country in the corresponding field. It is evident that Asian countries are particularly dominant in the number of articles mentioning AI within this field. In contrast to the ON field, countries with a high number of articles mentioning AI in the NN field tend to publish a greater proportion of Type C studies. These studies are more focused on advancing the field itself and only have an indirect connection to AI, suggesting that the high-ranking countries may exhibit stronger research capabilities within the NN field, rather than necessarily demonstrating superior expertise in leveraging AI for scientific research. However, it is worth noting that Singapore distinguishes itself with the highest proportion of articles mentioning AI in relation to its total articles in the NN field. This may be attributed to Singapore's advanced information technology infrastructure and its prioritization of artificial intelligence (Zahra et al., 2021), which has facilitated considerable activity both in using AI for NN research and in conducting NN research to advance AI.

(a) Type distribution of articles



(b) Statistics information

Country	Number	Share
China	696	0.37%
USA	221	0.27%
South Korea	130	0.37%
Singapore*	70	0.75%
India	57	0.25%
UK	48	0.25%
Germany	44	0.17%
Japan	41	0.20%
Canada	28	0.28%
Italy	25	0.21%

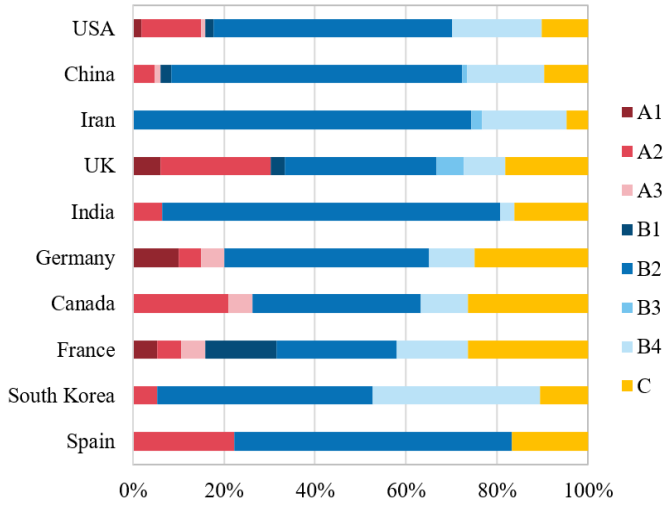
Figure 5. Number, distribution of types and share in all articles in the NN field of the largest contributing countries.

Note: * indicates that the proportion of articles mentioning AI in the total publications of the country is relatively prominent among all countries.

Figure 6 presents the distribution of types of articles mentioning AI published in the field of *Meteorology and Atmospheric Sciences* (MA) by the ten largest contributing countries, along with the proportion of articles mentioning AI within the total of articles from each country in the corresponding field. An intriguing finding is that

Iran has published 43 articles mentioning AI, accounting for 12% of all articles mentioning AI in this field. Moreover, the proportion of articles mentioning AI among Iran's total publications in this field is notably high (2%), significantly surpassing that of other countries. Furthermore, we conducted a search of Iran's publications across all disciplines and found that articles mentioning AI make up only 3% of its total output. These findings indicates that the integration of AI in Iran's MA research is exceptionally pronounced. Upon further examination of the thematic features of these studies, we found that they cover topics such as drought, flood and landslide prediction. However, the underlying factors contributing to Iran's exceptional performance in MA articles mentioning AI warrant further investigation. In addition, the contrast between the greater emphasis on "Science for AI" in Western countries and the focus on "AI for Science" in Asian countries is also evident in MA research.

(a) Type distribution of articles



(b) Statistics information

Country	Number	Share
USA	107	0.22%
China	83	0.21%
Iran*	43	2.04%
UK	33	0.21%
India	31	0.35%
Germany	20	0.14%
Canada	19	0.25%
France	19	0.21%
South Korea	19	0.38%
Spain	18	0.34%

Figure 6. Number, distribution of types and share in all articles in the MA field of the largest contributing countries.

Note: * indicates that the proportion of articles mentioning AI in the total publications of the country is relatively prominent among all countries.

Conclusions and discussion

This study examines the bidirectional relationship between AI and science, using the frequency and context of AI mentions in research articles as a source of information. It constructs a classification framework for the various ways in which AI is mentioned in research articles, with the aim of providing a quantitative approach to elucidating the role of AI in scientific research. This framework contributes to a clearer understanding of the interactive relationship between AI and science, revealing that AI emerges from the cross-disciplinary integration of knowledge and,

in turn, empowers and enhances research across different academic fields in diverse ways.

We conducted empirical research in three representative fields of research. The findings indicate that the mentions of AI in research articles becomes increasingly prevalent but also varies across countries with different levels of research capacity, geographic locations, and national contexts. These differences not only reflect the unique characteristics of each country but may also offer insights into potential collaborations between nations in the scientific discovery in the age of AI.

A limitation in our study is that it is confined to analyzing mentions to AI within research articles, which may not fully encompass the diverse ways in which AI contributes to scientific research and vice versa. However, the inclusion of AI mentions in research articles reflects the forms of AI engagement in scientific research that are widely recognized and accepted within the academic community, thereby providing significant indications of the relationship between AI and science. Another limitation arises from the search terms used to identify articles mentioning AI, which were based primarily on general representations and did not explore specific AI sub-technologies commonly employed in particular research fields. A more thorough investigation into the characteristics of individual fields would enable us to conduct a more comprehensive search for AI-related research within those specific domains. Moving forward, we aim to improve the classification framework by observing a broader range of samples from additional fields, and develop approaches to achieve automatic annotation. It will facilitate the exploration of more nuanced patterns over extended time periods, across diverse domains, and within larger datasets.

Acknowledgments

The authors would like to acknowledge support from the National Natural Science Foundation of China (Grant Nos. 72374160, L2424104) and the National Laboratory Centre for Library and Information Science at Wuhan University.

References

Arencibia-Jorge, R., Vega-Almeida, R. L., Jiménez-Andrade, J. L., & Carrillo-Calvet, H. (2022). Evolutionary stages and multidisciplinary nature of artificial intelligence research. *Scientometrics*, *127*(9), 5139-5158.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533-538.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'?. *Journal of Econometrics*, *65*(1), 83-108.

CB Insights. (2024). *Game Changers 2025: 9 technologies that will change the world*. https://www.cbinsights.com/reports/CB-Insights_Game-Changers-2025.pdf

European Commission: Directorate-General for Research and Innovation, Arranz, D., Bianchini, S., Di Girolamo, V., & Ravet, J. (2023). *Trends in the use of AI in science: a bibliometric analysis*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/418191

Gao, J., & Wang, D. (2024). Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nature Human Behaviour*, 1-12.

Google. (2024, November 18). 9 ways AI is advancing science. *The Keyword*. https://blog.google/technology/ai/google-ai-big-scientific-breakthroughs-2024/

Griffin, C., Wallace, D., Mateos-Garcia, J., Schieve, H., & Kohli, P. (2024). *A new golden age of discovery: Seizing the AI for Science opportunity*. Google DeepMind. https://deepmind.google/public-policy/ai-for-science/

He, Y. H. (2024). AI-driven research in pure mathematics and theoretical physics. *Nature Reviews Physics*, *6*(9), 546-553.

Ioannidis, Y. (2024). The 5th Paradigm: AI-Driven scientific discovery. *Communications of the ACM*, *67*(12), 5-5.

Frank, R. (1988). "Interdisciplinary": The first half century. *Issues in Integrative Studies*, *6*, 139-151.

Mariani, M., & Dwivedi, Y. K. (2024). Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research*, *175*, 114542.

McKinney, S. M., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89-94.

Messeri, L., Crockett, M.J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*, 49–58.

Miao, Q., & Wang, F. Y. (2024). AI4S Based on Parallel Intelligence. In *Artificial Intelligence for Science (AI4S) Frontiers and Perspectives Based on Parallel Intelligence* (pp. 1-19). Cham: Springer Nature Switzerland.

Miao, Q., Huang, M., Lv, Y., & Wang, F. Y. (2022, November). Parallel learning between science for AI and AI for science: a brief overview and perspective. In *2022 Australian & New Zealand Control Conference (ANZCC)* (pp. 171-175). IEEE.

OECD. (2023). *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*. OECD Publishing, Paris. https://doi.org/10.1787/a8d820bd-en

Royal Society. (2024). *Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research*. https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf

Szymanski, N. J., et al. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, *624*(7990), 86-91.

van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature*, *621*(7980), 672-675.

Vert, J. P. (2023). How will generative AI disrupt data science in drug discovery?. *Nature Biotechnology*, *41*(6), 750-751.

Wang, H., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47-60.

Xu, R., et al. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, *61*(3), 103665.

Zahra, A. A., & Nurmandi, A. (2021, March). The strategy of develop artificial intelligence in Singapore, United States, and United Kingdom. In *IOP Conference Series: Earth and Environmental Science* (Vol. 717, No. 1, p. 012012). IOP Publishing.

https://doi.org/10.51408/issi2025_047

Science and Artificial Intelligence: A Copyright Perspective

Dmitry Kochetkov

d.kochetkov@cwts.leidenuniv.nl, dmitry.kochetkov@urfu.ru
Centre for Science and Technology Studies, Leiden University, Kolffpad 1, 2333 BN Leiden (The Netherlands)
Ural Federal University, 19 Mira Street, 620062 Ekaterinburg (Russian Federation)

Abstract

Artificial intelligence (AI), particularly generative AI (GenAI) and large language models (LLMs), is transforming scientific research and higher education, offering new opportunities while raising significant ethical, legal, and regulatory challenges. This opinion piece explores the intersection of AI and science, focusing on the implications for copyright, peer review, and open science. AI systems, such as LLMs, are increasingly used in research applications, including text generation, data analysis, and peer review, with recent studies suggesting that AI-assisted reviews may improve efficiency and address reviewer shortages. However, concerns about bias, confidentiality, and the lack of guidelines for AI use in peer review persist. The rise of AI also poses challenges to copyright, as LLMs often rely on vast datasets of scientific works, raising questions about fair use, attribution, and licensing. Current regulatory frameworks in the United States, China, the European Union, and the United Kingdom focus on promoting innovation and responsible AI development, but gaps remain, particularly in addressing the use of copyrighted works for AI training. Creative Commons licenses, widely used for open-access outputs, do not fully address the complexities of AI training, and the absence of proper attribution in AI systems challenges the concept of originality. This paper calls for action to ensure that AI training is not considered a fair use exception to copyright law, advocating for authors' rights to refuse the use of their works for AI training and for universities to take a leading role in regulating AI. Governments and international organizations must develop harmonized legislative measures to protect authors' rights and ensure transparency in AI training datasets. The paper concludes that while AI offers transformative potential for science, a careful and responsible approach is needed to balance innovation with ethical and legal considerations, preventing the emergence of an oligopolistic market that prioritizes profit over scientific integrity.

Introduction

While there is no single, universally accepted definition of artificial intelligence (AI), it can be broadly defined as the ability of machines to learn, make decisions, and solve problems in a way that resembles human cognition (Sonone & Dharme, 2019). AI systems are designed to go beyond simple calculations, aiming to solve complex problems autonomously (Fogel, 2005). Generative AI (GenAI), a branch of AI, utilizes deep learning techniques – specifically generative models – to produce creative outputs such as music, images, and text (Ramdurai & Adhithya, 2023). In this opinion piece, I will focus primarily on Large Language Models (LLMs), which are intelligent systems capable of natural language processing (Gao et al., 2023; Hadi et al., 2023). These systems can process and generate human-like language, including tasks like machine translation. However, the true nature of their intelligence remains a subject of debate. Some researchers argue that the apparent intelligence of LLMs may reflect the interviewer's own intelligence rather than the model's, suggesting a "reverse Turing test" (Sejnowski, 2023).

The term "artificial intelligence" (AI) was first coined by John McCarthy at the Dartmouth Conference in 1956, marking the official beginning of AI's history (Strickland, 2021). However, the evolution of AI in the 20th century was marked by significant scientific and technical challenges that have hindered its rapid development. These challenges mainly include computational power limitations and algorithmic constraints (Puttgen & Jansen, 1987). Two periods, usually referred to as “AI winter,” represent the situation of reduction of interest and funding for AI research due to unmet expectations and the failure to deliver the promised breakthroughs. The first AI winter occurred in the 1970s and 1980s, primarily due to the overpromising of AI capabilities by researchers and the subsequent failure to achieve these goals. Similarly, the second AI winter in the late 1980s and early 1990s was caused by the failure of expert systems to deliver on their potential, despite significant investments by corporations (Lloyd, 1995). Algorithmic advances have played a crucial role in overcoming computational limitations (Selman, 2000).

Since then, the 21st century has witnessed significant advancements in AI, driven by increased computational power and the availability of vast amounts of data (Hwang, 2018; Liu et al., 2018). These advancements have transformed various sectors, including healthcare, finance, and manufacturing. AI's impact on society and the global order is profound, with implications extending far beyond technology (Rama Padmaja & Lakshminarayana, 2024). The rise of AI has reshaped power dynamics among nations, with countries like the USA, China, and Russia leading the global race for AI dominance (Vijayakumar, 2023).

The development of AI technology presents both challenges and opportunities across various fields (Rama Padmaja & Lakshminarayana, 2024; Wolff et al., 2018). While AI offers immense potential, its advancement raises ethical concerns, including biases, privacy issues, and broader social implications (Rama Padmaja & Lakshminarayana, 2024). Li (2023) identifies 12 key ethical concerns and related strategies for applying AI in healthcare: justice and fairness, freedom and autonomy, privacy, transparency, patient safety and cybersecurity, trust, beneficence, responsibility, solidarity, sustainability, dignity, and conflicts. AI's influence spans all five dimensions of sustainability, with both positive and negative consequences (Khakurel et al., 2018). For instance, an analysis of a Google Scholar sample of questionable scientific papers suspected to be generated by GPT revealed that many address applied, often controversial issues prone to misinformation, such as environment, health, and computing (Haider et al., 2024). Additionally, LLMs may pose a threat to copyright, as they can generate content that potentially violates intellectual property rights (German, 2024). Currently, neither copyright nor “open” licenses can protect scholarly content from unauthorized reuse in AI training (Decker, 2025).

AI is transforming research jobs, and science, that in turn provides LLMs with a vast amount of data for training. The goal of this opinion piece is to analyze the potential consequences of the further development of AI on science, highlighting its positive effects while also mitigating risks. In the next section, I will provide a brief overview of how AI is being used in research applications. I will then analyze the current state of AI regulation, particularly regarding science, identifying any gaps in the current

regulations. Finally, I will outline several suggestions for filling these gaps to ensure the safe and effective use of AI in academic research.

Applications of AI and LLMs in Research and Higher Education

Artificial intelligence (AI), particularly large language models (LLMs), is transforming higher education and research in much the same way it is revolutionizing other industries. AI has the potential to enhance personalized learning experiences, provide feedback to students, identify at-risk learners, and accelerate the research process (Tarisayi, 2024). Applications of AI in these fields include text generation, data analysis, literature review assistance, and peer review (Alqahtani et al., 2023). For instance, AI can automate many tasks involved in conducting systematic literature reviews (De La Torre-López et al., 2023). Another promising use case is the proofreading and editing of scientific texts. While these applications have the potential to revolutionize education and research, challenges remain, including ethical concerns, algorithmic bias, and the need for human oversight (Alqahtani et al., 2023; Peláez-Sánchez et al., 2024). Algorithmic bias refers to systematic errors in AI systems that can lead to unfair and unequal outcomes (Shin & Shin, 2023). Furthermore, Andersen et al. (2024) identified three clusters of AI perception among academics: "GenAI as a workhorse," "GenAI as a language assistant only," and "GenAI as a research accelerator." The authors argue that these variations reflect differences across disciplines and knowledge production models.

Automatic or AI-assisted peer review has been proposed as a potential solution to issues of quality and reproducibility in scientific research. Software tools for automatically evaluating scientific papers using AI, StatReviewer[1] and UNSILO[2], have emerged in recent years[3]. Additionally, tools like the *statcheck* package for verifying statistical analyses have gained traction[4]. Until recently, these tools were considered auxiliary and incapable of replacing human labor (Baker, 2015; Heaven, 2018). However, recent advances in AI are challenging this notion.

Recent studies have explored the impact of AI and LLMs on peer review, with research indicating that AI-assisted reviews are becoming more prevalent. At ICLR 2024, it is estimated that at least 15.8% of reviews will be AI-assisted (Latona et al., 2024). These AI-assisted reviews tend to assign higher scores to papers and increase acceptance rates (Latona et al., 2024), potentially improving review quality and addressing reviewer shortages (Hosseini & Horbach, 2023). However, such studies are often based on limited samples. For example, Biswas et al. (2023) compared ChatGPT's performance as an AI reviewer to human reviews for a single published article. The authors found that ChatGPT demonstrated commendable ability in identifying methodological flaws, providing insightful feedback on theoretical

[1] StatReviewer. URL: http://statreviewer.com/ (date of access: 22.01.2024).
[2] UNSILO. URL: https://site.unsilo.com/site/ (date of access: 22.01.2024).
[3] At the same time, plagiarism detection systems have existed for much longer. For example, "Antiplagiat," a well-known system in Russia, was established in 2005.
[4] statcheck. URL: https://michelenuijten.shinyapps.io/statcheck-web/ (date of access: 22.01.2024), also R package.

frameworks, and assessing the overall contribution of articles to their respective fields.

Despite these advancements, concerns about bias amplification, confidentiality, and the lack of guidelines for LLM use in peer review persist (Hosseini & Horbach, 2023). Some researchers advocate for AI to assist with manuscript triaging (Bauchner & Rivara, 2024), suggesting that human-AI collaboration could democratize academic culture (Sarker et al., 2024). Nevertheless, researchers recommend disclosing the use of LLMs and maintaining human responsibility for review accuracy and integrity (Hosseini & Horbach, 2023).

The impact of AI on the publishing industry can be described as revolutionary. It is expected that AI will bring about a third digital transformation in the industry (Bergstrom & Ruediger, 2024). Two possible scenarios for the future development of AI in scholarly publishing have been proposed. In the first scenario, AI would make the publishing process more efficient, expanding the range of services offered by publishers. In a more radical scenario, AI could fundamentally change the way scientific communication occurs, transforming the channels used for communication.

The interaction between generative AI (GenAI) and the open access movement is complex (Hosseini et al., 2024). GenAI can make scholarly publications more comprehensible to the public or researchers from other fields. It can also help mitigate the negative consequences of information overload and assist researchers in fully benefiting from open access. However, significant risks are associated with using GenAI to enhance access to scholarly literature. One concern is the potential for systems to provide inaccurate or biased summaries, syntheses, or advice. Another risk is the possibility of facilitating the proliferation of paper mills. Finally, the absence of proper attribution of training data challenges the concept of originality and may discourage the sharing of data and papers.

Open science has led to the generation of vast amounts of data, presenting both opportunities and challenges for the scientific community. AI research can also be part of open science, particularly through the development of open-source LLMs such as Game 2, Nemo Tron-4, and Llama 3.1. Open datasets are crucial to the success of these open-source projects. However, developers face numerous challenges, including language bias and safety issues.

Several community initiatives aim to address these challenges. One such initiative is the Aya project, which seeks to bridge the language barrier by providing a human-curated instruction-following dataset in 65 different languages (Singh et al., 2024). The dataset contains 513 million examples across 114 languages. As a result of this initiative, three key resources have been developed and made freely available: the Aya Dataset, the Aya Collection, and the Aya Evaluation Suite. This initiative serves as a platform for future research collaboration to continue bridging the gap in language resources.

Another issue with open-source LLMs is their susceptibility to malicious exploitation. Yi et al. (2024) identified vulnerabilities in the safety alignment of open-access LLMs, which can significantly increase the success rate and

harmfulness of jailbreak attacks[5]. The study proposes two types of techniques that can make LLMs adeptly reverse-aligned to output harmful content, even in the absence of manually curated malicious datasets.

AI-Related Regulations

In this section, I provide a brief analysis of the regulations related to artificial intelligence (AI) in the United States, China, the United Kingdom, and the European Union.

Interestingly, there is currently no comprehensive regulation governing AI in the UK and the US. The Sunak government issued a framework document in 2023 titled *A Pro-Innovation Approach to AI Regulation* (Department for Science, Innovation & Technology, 2023), which establishes basic principles for AI. The document promotes flexible regulation and aims to foster innovation through the development and use of AI technologies. The British government has also expressed its ambition to make the UK the best place to invest in AI.

In the United States, a framework document was published in October 2023, titled Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023). Notably, this document includes actions related to copyright law, stating: "…consult with the Director of the United States Copyright Office and issue recommendations to the President on potential executive actions relating to copyright and AI. The recommendations shall address any copyright and related issues discussed in the United States Copyright Office's study, including the scope of protection for works produced using AI and the treatment of copyrighted works in AI training."

A significant step forward was taken with the development of the *Generative AI Copyright Disclosure Act of 2024* (H.R.7913 - 118th Congress, 2023-2024). This act aims to ensure transparency in the use of copyrighted works for AI training and is currently under consideration in the House of Representatives. If passed, the act would require companies to notify the U.S. Copyright Office about any copyrighted works used in their AI systems. These notifications must be submitted 30 days before or after the public release of the AI system, ensuring transparency and accountability. The act is intended to help copyright holders make informed decisions about licensing and compensation. However, the wording of the document remains vague, raising questions for both AI developers and copyright owners. Additionally, I have concerns about the inability of copyright holders to prohibit the use of their works for AI training, which creates a bias in favor of AI development.

In China, the *Interim Measures for the Management of Generative Artificial Intelligence Services* (Cyberspace Administration of China et al., 2023) were implemented on August 15, 2023. These regulations, comprising 24 articles, aim to strike a balance between fostering innovation and ensuring the security and governance of AI. Article 3 emphasizes the importance of maintaining a harmonious

[5] User prompt injection attacks occur when users deliberately exploit system vulnerabilities to elicit unauthorized behavior from an LLM (see, for example, https://learn.microsoft.com/en-au/azure/ai-services/content-safety/concepts/jailbreak-detection).

relationship between development and innovation while prioritizing security and governance in the field of AI. Articles 5 and 6 highlight the need for collaboration in developing basic technologies, such as chips and software platforms, as well as the creation of shared data resources. Article 16 states that all regulatory measures must be compatible with innovation, and Article 2 clarifies that the regulations apply only to publicly available generative AI services. Service providers are held responsible for the content generated using their services. Chinese regulations are among the most stringent in the world. For example, Article 12 mandates that users must be informed when content is generated using AI as a blanket rule.

On August 1, 2024, the European *Artificial Intelligence Act* (AI Act) entered into force (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024). This law primarily aims to reduce the risks associated with the use of AI. It focuses mainly on high-risk AI systems, while low-risk systems require transparency. For instance, chatbots must clearly inform users that they are interacting with a machine, and certain AI-generated content must be labeled as such.

In summary, the legislative frameworks in major AI-developing countries primarily focus on either the responsible development and use of AI[6], including content labeling, or on fostering innovation and attracting investment in the AI industry. Notably, only the US addresses copyright issues in connection with AI development, but its regulatory framework remains incomplete and appears biased toward AI developers rather than copyright holders. It is also worth noting that, at present, there are no specific legislative regulations governing AI in the Russian Federation. However, Russia has introduced the concept of "experimental legal regimes for digital innovations[7]," which allows for the testing of technologies that are not yet legally regulated.

Copyright and Licensing

Most scientific works are protected by copyright laws. Copying and retaining these works in AI systems, as well as reproducing them in outputs, involves copyright, making appropriate licensing essential for compliance (Johnson, 2024). The generated output can be considered a derivative work, although this is not explicitly stated in any legal documents.

Creative Commons (CC) licenses are the most widely used for open-access outputs. Approximately 28% of global research output is licensed under the Creative Commons Attribution license (CC BY), while another 22% uses more restrictive Creative Commons licenses (Pollock & Michael, 2024). However, Creative Commons acknowledges that existing CC licenses do not fully address the specific challenges related to using creative works for AI training (Walsh, 2023). Using CC-licensed content raises several questions, such as whether the attribution requirement is fulfilled when training LLM models. In my opinion, this is not the case. For example, the training dataset for ChatGPT contains millions of scientific articles, but

[6] Living Guidelines on the Responsible Use of Generative AI in Research | Research and Innovation (2024) also focuses on responsible use of AI and related issues of research integrity.

[7] Regulated by Federal Law No. 258-FZ, dated July 31, 2020, "On Experimental Legal Regimes in the Field of Digital Innovations in the Russian Federation".

it is unclear exactly which ones were used ("AI Firms Must Play Fair When They Use Academic Data in Training," 2024).

However, if the use of content is subject to copyright exclusions, the licensee's abilities are limited. In fact, such an exclusion is currently being considered for legislation in the US. Moreover, the US fair use doctrine allows for the unlicensed use of copyrighted works under certain circumstances. AI training is often considered a case of fair use (Johnson, 2024; Walsh, 2023). For instance, OpenAI argues that this position is "supported by long-standing and widely accepted precedents" (*OpenAI and Journalism*, 2024).

Publishers are also responding to market changes by developing licensing agreements for the use of content in LLM training (Schonfeld, 2024). Currently, the number of such deals is relatively low[8], and they primarily cover content distributed through subscription services. If a publishing contract includes the full transfer of rights to the publisher, the publisher can license the content for AI training without seeking the authors' consent (Hansen, 2024). This underscores the importance of the rights retention strategy. Major publishers, along with Clarivate, are rapidly developing new AI-based businesses, which are evolving into data cartels (Pooley, 2024). This could lead to a situation where the academic AI market adopts the same oligopolistic structure as the current academic publishing market.

A Call for Action

Science and artificial intelligence (AI) are closely linked. Research provides data, which is crucial for training large language models (LLMs) and advancing data science more broadly. At the same time, generative AI (GenAI) is revolutionizing research. Open-source LLMs are an essential part of open science. While AI presents significant opportunities for scientific advancement, it also poses substantial risks. Legislation in this field is still evolving, and regulatory and policy documents often focus on attracting investment in AI or promoting its responsible development and use. The use of publicly available research outputs for training LLMs falls into a "grey area." At the moment, the community lacks any meaningful discussion on the reuse of academic content for LLMs' training. Attempts to raise this issue are made, but their impact is rather limited (Decker, 2025). Below, I offer some thoughts on actions that can be taken in the near future.

First and foremost, AI training should not be considered an exception to copyright law (i.e., under the fair use doctrine). Recognizing LLM training as a case of fair use undermines efforts to reform copyright regulation. In my opinion, LLM training should not qualify as fair use for at least two main reasons:

1. *Non-commercial use is not guaranteed:* Many AI systems already operate on paid subscription models. Even if no fees are currently charged, there are no legal restrictions preventing these models from becoming commercialized in the future.

[8] Generative AI Licensing Agreement Tracker. URL: https://sr.ithaka.org/our-work/generative-ai-licensing-agreement-tracker/.

2. *Content can be reproduced with high accuracy:* AI-generated content often closely resembles the original, making it subject to copyright and attribution requirements.

This issue is particularly relevant in the US context, but given that most AI developers are based in the US, it is critical for the global development of the industry. Some researchers argue that it will take years for US courts to address the issue of licensing content for LLM training (Bergstrom & Ruediger, 2024). This is a major concern for the academic community, as the market will continue to evolve, researchers will increasingly rely on AI for interacting with scholarly output, and it will become more difficult to implement changes (see below for further discussion of limitations and challenges).

Authors should have the option to refuse the use of their work for training GenAI models or specific groups of such models. This should be explicitly stated in the licensing terms. There are two possible strategies to achieve this:

1. *Examine existing licenses:* The Creative Commons BY-ND (Attribution-NoDerivatives) license could be considered restrictive for AI training, but only if regulatory frameworks recognize AI-generated content as derivative works. However, determining whether AI-generated content qualifies as a derivative work is complicated by the fact that LLMs can produce different responses for each query, making it difficult to assess similarity to the original. The BY-NC (Attribution-NonCommercial) license may also be restrictive for training models intended for commercial use[9].
2. *Introduce a new "NT" (no train) extension:* This would explicitly prohibit the use of licensed works for AI training. However, since the original datasets used for LLM training are not publicly accessible, the prospects for enforcing such licensing terms remain uncertain. Additionally, publishing contracts should specify that publishers cannot use articles to train their LLMs or other AI models without author consent.

Universities as Key Players in AI Regulation

Universities should take a leading role in regulating AI. On the one hand, universities often act as publishers or maintain their own repositories, making it feasible to implement content licensing approaches in practice. On the other hand, universities conduct research and develop GenAI models, placing them at the forefront of addressing the ethical aspects of these processes. Furthermore, universities can provide evidence to support legislative regulation. Having said that, I must acknowledge that universities lack the regulatory power that governments possess. However, it is concerning that many current community documents in the field of open science, such as the *Barcelona Declaration on Open Research Information* (2024), do not address AI-related issues.

[9] However, can we be certain that today's open models will not be commercialized in the future?

Legislative Measures and International Cooperation

Governments and international organizations must develop and implement legislative measures to protect authors' rights and prevent the unauthorized use of their works for training GenAI models. One of the first steps should be the mandatory disclosure of training datasets by developers.

The challenge lies not only in adopting national AI laws but also in harmonizing these laws globally. Without international coordination, commercial developers could exploit "safe harbors" to serve their own interests. Therefore, it is essential for large intergovernmental organizations, such as UNESCO, to take on this task. Another challenge is that AI models cannot be "untrained." If restrictions are imposed only on new models, existing models would gain a non-market advantage. Conversely, applying restrictions retroactively to existing models could destabilize the industry. A responsible dialogue is needed to find a balanced solution. One possible approach is retrieval-augmented generation, which allows models to reference relevant papers in their outputs ("AI Firms Must Play Fair When They Use Academic Data in Training," 2024).

Conclusion

The author of this article does not oppose AI. In fact, while writing this manuscript, the Yandex. Translate service was used to assist with reading Chinese text and proofreading the English version. The development of AI brings numerous opportunities for research, but it requires a careful and responsible approach that considers the interests of all stakeholders. Otherwise, there is a risk of fostering an oligopolistic market driven by profit maximization, resembling the current dynamics of the academic publishing sector. As an author, I would like the option to refuse the use of my work for training GenAI models, especially for commercial purposes.

Acknowledgments

The author would like to express his gratitude to Ludo Waltman and Erna Sattler for their valuable comments and suggestions, which have greatly improved this paper.

Contribution

Dmitry Kochetkov: Conceptualization, Writing – original draft

References

AI firms must play fair when they use academic data in training. (2024). *Nature*, 632(8027), 953–953. https://doi.org/10.1038/d41586-024-02757-z

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, 19(8), 1236–1242. https://doi.org/10.1016/j.sapharm.2023.05.016

Andersen, J. P., Degn, L., Fishberg, R., Graversen, E. K., Horbach, S. P. J. M., Schmidt, E. K., Schneider, J. W., & Sørensen, M. P. (2024). *Generative Artificial Intelligence*

(GenAI) in the research process – a survey of researchers' practices and perceptions. SocArXiv. https://doi.org/10.31235/osf.io/83whe

Baker, M. (2015). Smart software spots statistical errors in psychology papers. *Nature*. https://doi.org/10.1038/nature.2015.18657

Kramer, B., Neylon, C., & Waltman, L. (2024). *Barcelona Declaration on Open Research Information*. https://doi.org/10.5281/ZENODO.10958522

Bauchner, H., & Rivara, F. P. (2024). Use of artificial intelligence and the future of peer review. *Health Affairs Scholar*, 2(5), qxae058. https://doi.org/10.1093/haschl/qxae058

Bergstrom, T., & Ruediger, D. (2024). *A Third Transformation? Generative AI and Scholarly Publishing*. Ithaka S+R. https://doi.org/10.18665/sr.321519

Biswas, S., Dobaria, D., & Cohen, H. L. (2023). ChatGPT and the Future of Journal Reviews: A FeasibilityStudy. *The Yale Journal of Biology and Medicine*, 96(3), 415–420. https://doi.org/10.59249/SKDH9286

De La Torre-López, J., Ramírez, A., & Romero, J. R. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10), 2171–2194. https://doi.org/10.1007/s00607-023-01181-x

Decker, S. (2025, April 15). *Guest Post - The Open Access – AI Conundrum: Does Free to Read Mean Free to Train?* The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2025/04/15/guest-post-the-open-access-ai-conundrum-does-free-to-read-mean-free-to-train/

Department for Science, Innovation & Technology (2023). *A pro-innovation approach to AI regulation* (No. 815). https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Fogel, D. B. (2005). *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence* (1st ed.). Wiley. https://doi.org/10.1002/0471749214

Gao, Y., Baptista-Hon, D. T., & Zhang, K. (2023). The inevitable transformation of medicine and research by large language models: The possibilities and pitfalls. *MedComm – Future Medicine*, 2(2), e49. https://doi.org/10.1002/mef2.49

Generative AI Copyright Disclosure Act of 2024, H.R.7913—118th Congress (2023-2024) (2024). https://www.congress.gov/bill/118th-congress/house-bill/7913

German, D. M. (2024). *Copyright related risks in the creation and use of ML/AI systems* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.01560

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. https://doi.org/10.36227/techrxiv.23589741.v4

Haider, J., Söderström, K. R., Ekström, B., & Rödl, M. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-156

Hansen, D. (2024, July 30). *What happens when your publisher licenses your work for AI training?* Authors Alliance. https://www.authorsalliance.org/2024/07/30/what-happens-when-your-publisher-licenses-your-work-for-ai-training/

Heaven, D. (2018). AI peer reviewers unleashed to ease publishing grind. *Nature*, 563(7733), 609–610. https://doi.org/10.1038/d41586-018-07245-9

Hosseini, M., & Horbach, S. P. J. M. (2023). *Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other Large Language Models in scholarly peer review.* https://doi.org/10.21203/rs.3.rs-2587766/v1

Hosseini, M., Horbach, S. P. J. M., Holmes, K., & Ross-Hellauer, T. (2024). Open Science at the generative AI turn: An exploratory analysis of challenges and opportunities. *Quantitative Science Studies*, 1–24. https://doi.org/10.1162/qss_a_00337

Cyberspace Administration of China, National Development and Reform Commission of the People's Republic of China, Ministry of Education of the People's Republic of China, Ministry of Science and Technology of the People's Republic of China, Ministry of Industry and Information Technology of the People's Republic of China, Ministry of Public Security of the People's Republic of China, State Administration of Radio and Television (2023). *Interim Measures for the Management of Generative Artificial Intelligence Services*. http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

Johnson, B. (2024, July 31). *True or False? Addressing Common Assumptions About Copyright and AI*. Copyright Clearance Center. https://www.copyright.com/blog/addressing-common-assumptions-copyright-ai/

Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A., & Zhang, W. (2018). The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies*, 6(4), 100. https://doi.org/10.3390/technologies6040100

Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., & West, R. (2024). *The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.02150

Li, Y. (2023). Specifics of regulatory and legal regulation of Generative Artificial Intelligence in the UK, USA, EU and China. *Law Journal of the Higher School of Economics*, 3, 245–267. https://doi.org/10.17323/2072-8166.2023.3.245.267

Living guidelines on the responsible use of generative AI in research | Research and innovation. (2024). https://research-and-innovation.ec.europa.eu/document/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en

Lloyd, J. W. (1995). Surviving the AI Winter. *Logic Programming: The 1995 International Symposium*, 33–47.

OpenAI and journalism. (2024, January 8). https://openai.com/index/openai-and-journalism/

Peláez-Sánchez, I. C., Velarde-Camaqui, D., & Glasserman-Morales, L. D. (2024). The impact of large language models on higher education: Exploring the connection between AI and Education 4.0. *Frontiers in Education*, 9, 1392091. https://doi.org/10.3389/feduc.2024.1392091

Pollock, D., & Michael, A. (2024, December 10). *News and Views: How much content can AI legally exploit?* https://www.deltathink.com/news-and-views-how-much-content-can-ai-legally-exploit

Pooley, J. (2024). Large Language Publishing: The Scholarly Publishing Oligopoly's Bet on AI. *KULA: Knowledge Creation, Dissemination, and Preservation Studies*, 7(1), 1–11. https://doi.org/10.18357/kula.291

Puttgen, H., & Jansen, J. (1987). Knowledge based systems applied to power systems: A passing fad or a useful tool here to stay? *26th IEEE Conference on Decision and Control*, 408–412. https://doi.org/10.1109/CDC.1987.272830

Rama Padmaja, C. V., & Lakshminarayana, S. (2024). The rise of AI: A comprehensive research review. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13(2), 2226. https://doi.org/10.11591/ijai.v13.i2.pp2226-2235

Ramdurai, B., & Adhithya, P. (2023). The impact, advancements and applications of Generative AI. *International Journal of Computer Science and Engineering*, 10(6), 1–8. https://doi.org/10.14445/23488387/IJCSE-V10I6P101

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng

Sarker, S., University of Virginia, Susarla, A., Michigan State University, Gopal, R., University of Warwick, Thatcher, J. B., & University of Colorado / University of Manchester. (2024). Democratizing Knowledge Creation Through Human-AI Collaboration in Academic Peer Review. *Journal of the Association for Information Systems*, 25(1), 158–171. https://doi.org/10.17705/1jais.00872

Schonfeld, R. C. (2024, October 15). *Tracking the Licensing of Scholarly Content to LLMs.* The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2024/10/15/licensing-scholarly-content-llms/

Sejnowski, T. J. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation*, 35(3), 309–342. https://doi.org/10.1162/neco_a_01563

Selman, B. (2000). Compute-intensive methods in artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, *28*(1/4), 35–38. https://doi.org/10.1023/A:1018943920174

Shin, D., & Shin, E. Y. (2023). Data's Impact on Algorithmic Bias. *Computer*, 56(6), 90–94. https://doi.org/10.1109/MC.2023.3262909

Singh, S., Vargus, F., D'souza, D., Karlsson, B., Mahendiran, A., Ko, W.-Y., Shandilya, H., Patel, J., Mataciunas, D., O'Mahony, L., Zhang, M., Hettiarachchi, R., Wilson, J., Machado, M., Moura, L., Krzemiński, D., Fadaei, H., Ergun, I., Okoh, I., … Hooker, S. (2024). Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 11521–11567. https://doi.org/10.18653/v1/2024.acl-long.620

Sonone, S., & Dharme, A. (2019). A review paper on simulated intellect. *International Journal of Physics and Mathematics*, 1(1), 31–33. https://doi.org/10.33545/26648636.2019.v1.i1a.6

Strickland, E. (2021). The Turbulent Past and Uncertain Future of AI: Is there a way out of AI's boom-and-bust cycle? *IEEE Spectrum*, 58(10), 26–31. https://doi.org/10.1109/MSPEC.2021.9563956

Tarisayi, K. S. (2024). Strategic leadership for responsible artificial intelligence adoption in higher education. *CTE Workshop Proceedings*, 11, 4–14. https://doi.org/10.55056/cte.616

Vijayakumar, A. (2023). Potential impact of artificial intelligence on the emerging world order. *F1000Research*, 11, 1186. https://doi.org/10.12688/f1000research.124906.2

Walsh, K. (2023, August 18). *Understanding CC Licenses and Generative AI. Creative Commons*. https://creativecommons.org/2023/08/18/understanding-cc-licenses-and-generative-ai/

Wolff, J., Gordon, S., & Guo, D. (2018). The Rise of Artificial Intelligence. *Advances in Social Sciences Research Journal*. https://doi.org/10.14738/assrj.56.4722

Yi, J., Ye, R., Chen, Q., Zhu, B., Chen, S., Lian, D., Sun, G., Xie, X., & Wu, F. (2024). On the Vulnerability of Safety Alignment in Open-Access LLMs. *Findings of the Association for Computational Linguistics ACL 2024*, 9236–9260. https://doi.org/10.18653/v1/2024.findings-acl.549

https://doi.org/10.51408/issi2025_031

Science-Policy Tendencies in Armenia Towards the International Collaboration

Gevorg Kesoyan[1], Aram Mirzoyan[2], Simon Hunanyan[3], Miranush Kesoyan[4]

[1]*gevorgkesoyaned@gmail.com*, [2]*aram.mirzoyan@asnet.am*,
[3]*simhunanyan@gmail.com*, [4]*mkesoyan1996@gmail.com*
Center for Scientific Information Monitoring and Analysis (CSIAM), Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia (Republic of Armenia)

Abstract

Since its independence (1991) the Republic of Armenia has faced the challenge of preserving and further development of science and technology. In this regard the role of international scientific cooperation is especially emphasized. The article explores the state policy in this field with the focus on the research of the bilateral.

Methodologically, the study relies on the principles of scientometric analysis. The methods include the desk research, quantitative measurements and data retrieval from Web of Science. The research data were retrieved from the interstate agreements, bilateral competitions, and from Web of Science. The research results showed that 102 cooperation agreements have been signed with CIS and EU member states and countries from Asia and America. By maintaining cooperative relations with the countries of the former Soviet Union, scientific cooperation with European and other countries is also developing. On their basis, bilateral international competitions have been organized since 2009. These competitions secured funding for 332 projects. There are 377 publications[1] linked with the winning programs in the WOS database. Among the funded programs and the publications within their scope, the dominant part is from natural sciences, mainly from physics. The publications were excellently made with co-authorship, which contributed to the development of international cooperation. These results of the study contribute to the creation of the picture of international scientific collaboration in the frames of the bilateral competitions and can help to make the necessary refinements.

Introduction

As a result of the collapse of the Soviet Union at the end of December 1991, the former Soviet Republics, that became independent, found themselves in the face of new challenges and opportunities in the field of international relations. The right to conduct an independent policy and the elimination of the "Iron Curtain" made it possible to establish relations not only with the countries of the former Eastern bloc, but also with many Western and Asian states. In addition to establishing interstate diplomatic, economic and political ties, the Republic of Armenia also pays great attention to the formation of a new network of international scientific cooperation. Armenian scientists faced the following dichotomous options: a) collaboration with the former Soviet states, with which they had long history and tradition of scientific collaboration; b) collaboration with European countries, but it was in many ways a terra incognita. (Sargsyan et al.) The latter option could be broadened and included USA and Canada as well as South Eastern countries. The reason for that was not merely the desire to discover new horizons which were often unavailable under the

[1] As of June 22, 2022.

Soviet rule, but the imperative. The first collaborative article was published in 1665 (Gazni & Didegah, 2011) and since then the collaborative publications started to gain more and more popularity in scientific research. This process reflected the growing importance of scientific collaboration. And since many decades international scientific collaboration is a common feature in scientific research. (Coccia & Wang, 2015) More precisely the real heyday of the international scientific cooperation has been started since the second half of the 20th century. (Astakhova, 2020) As the result it is possible to argue that we are witnessing the era of international collaboration in the field of scientific research. (Gui et al., 2019) International scientific collaboration has shown the steady growth in all research fields. (Coccia & Wang, 2016) The exchange of scientific knowledge and skills make the shift of the focus of science from the national to the global level. (Gazni et al., 2012) Adams claims that now we are witnessing the forth age of research – which is driven by scientific collaboration. (Adams, 2013) He states that the first three ages were the individual, the institutional and the national. The different reasons can be mentioned for the development of the international collaboration: very often the breakthrough research projects are too complicated in order to be conducted in a single state (Astakhova, 2020); scientific collaboration has a proven positive effect on the scientific as well as on the economic productivity (Pfotenhauer et al., 2016); researchers' wish to increase their scientific popularity, visibility and recognition; the growing need of the rationalization of scientific manpower; increasing specialization in science; continuously growing amount of knowledge that needs to be put together in order to have significant advances in science; the phenomenon of cross-fertilization across disciplines (Katz & Martin, 1997); seeking excellence of the research, increase of the visibility of the research which can be resulted of the higher citation rank, capacity building (The Royal Society, 2011) etc.

It is important to mention that research collaboration can take place on three levels – micro (collaboration between individual scientists), meso (collaboration between organizations) and macro (collaboration between countries. Looking ahead it must be said that in case of Armenia all three types of scientific collaborations are available. Scientific collaboration has stronger impact in the "hard" sciences, than in the "soft" ones. (Bote et al., 2012) When speaking about the scientific collaboration and especially the co-authorship it should be kept in mind that sometimes the collaborative papers can be just the "mandatory exercises" in the frames of bilateral agreements on the different levels. (Glänzel, 2001) But in general the growth of scientific output during the last decades is provided mainly due to the international scientific collaboration. It is especially true for the Western countries. (Adams, 2013) The research shows that in the Western hemisphere the number of domestic papers (publications that have authors only from home country) does not go through any visible changes and is stable.

In the Republic of Armenia, the development of science and technology has been declared a priority and an important place is given to the development of international scientific cooperation. As stated by Finardi & Buratti (2016) "International scientific collaboration is strategic for the growth of a country, in particular for developing countries". But it is not a one way process. The

international scientific cooperation can be considered as one of the components of the process of globalization of science. And the latter is a kind of a "win-win" game when both advanced and developing countries benefit. (Freeman, 2010) Moreover there is an approach according to which the best science comes from international collaboration. (Coccia & Wang, 2015) It should be also added that co-authored publications are usually cited more than single author ones and that internationally co-authored papers are also usually cited more than single country ones. (Sooryamoorthy, 2009) The reason for it is the larger potential community. (Schmoch & Schubert, 2008) So it can be stated that the co-authorship increases the papers' impact.

According to a number of researchers, international scientific cooperation becomes possible when certain principles coincide. Indian researcher Nagpaul (2003) believes that geographic, thematic, sociocultural priorities are of fundamental importance for creating a network of scientific cooperation. Schott (1991) concluded that international cooperation depends on "political, cultural and social factors". Moëd et al. believe that "The differences between countries with respect to international scientific integration are affected by both the policies of the national governments and long-term traditions in the political, economic and cultural fields". (Moëd et al., 1991, p. 308) On the one hand, the Republic of Armenia has maintained and developed scientific cooperation with the states of the former USSR; on the other hand, it has found points of intersection of interests that have made it possible to establish cooperation with dozens of states in Europe, Asia and America in the field of science.

Cooperation developed on three levels. At the first level, the Armenian government and the authorized state body for science have signed cooperation agreements with dozens of foreign countries. In addition to cooperation with the states of the former Soviet Union, the importance of cooperation with Western states was emphasized at the state level. At the second level, academic institutions, universities and other scientific organizations of Armenia have established cooperation with relevant foreign structures. The third level is the individual one: researchers from Armenia collaborate with their foreign colleagues, resulting in thousands of joint scientific papers. As the result of it, there is an interaction and mutual influence of different cultures, mutual cognition and localization of international scientific achievements. According to Gomez et al. (1999) "Globalization of science reflects itself in an increasing cooperation between nations which originates different types of scientific collaboration networks, frequently enhanced by science policy measures taken at national and supranational levels".

State research grants (as well as delivered from private sector/industry) have significant impact on seeding and fostering fundamental and cutting-edge research projects, which leads to research innovations and scientific discoveries. (Wang et al., 2020) Bilateral competitions can be considered as the part of the science diplomacy. The latter can be described using the words of Nina Fedoroff, once science and technology adviser to the US Secretary of State: "science diplomacy is the use of scientific collaborations among nations to address the common problems facing

twenty-first century humanity and to build constructive international partnerships." (Ruffini, 2018, 11-12)

In this article, we aim to find out the activities carried out by the Republic of Armenia in the direction of establishing international cooperation in the field of science at the level of agreements reached with foreign countries and conducted competitions, the opportunities created by competitions in the development and internationalization of various fields of science. To achieve this goal, the following issues were discussed:

1. With which countries have cooperation agreements been signed and with which countries is cooperation more active and developing?
2. Which part of signed international agreements led to practical work in the context of organizing bilateral competitions?
3. Which specialties granted the opportunity to participate in the competitions and which specialties met the requirements of the competitions?
4. The volume of international articles published within the framework of the winning programs, their distribution by fields. What part of those articles is the result of international cooperation?

Previously the Center for Scientific Information Analysis and Monitoring has already conducted the research concerning the collaboration of Armenian and Russian scientist in the frames of the Russian-Armenian bilateral competitions which resulted with publishing of two articles that presented the role of such competitions in the promotion of scientific collaboration between Armenia and Russia (Gzoyan et al., 2017) collaboration of the Armenian and Russian scientists in the frames of bilateral competitions (Glukhov et al., 2017). This article presents the logical continuation of the aforementioned research and deals with its whole specter.

Data and method

This work is based on the international documents[2] signed by the scientific policy makers of the Republic of Armenia - the Government[3] and the authorized state body in the field of science,[4] joint international competitions[5] held on their basis and their results.[6]. On the basis of this information we have created 3 databases: documents on scientific cooperation, announced bilateral competitions and winning projects of competitions. The information for the analysis of the articles published in the winning projects was extracted from the international scientific information database Web of Science. The time frames include the entire period of independence of the Republic of Armenia, starting from 1991 until the first half of 2022.

[2] We have considered agreements, contracts, memoranda, programs, which we used in scientometric calculations as documents of equal force.

[3] Legal information system of Armenia, https://www.arlis.am/

[4] Science Committee of Ministry of Education, Science, Culture and Sports RA, http://scs.am/am/0652fc7e4429cb2579571955.

[5] Science Committee of Ministry of Education, Science, Culture and Sports RA, http://scs.am/am/ef52f2239b1bc62940173436

[6] Science Committee of Ministry of Education, Science, Culture and Sports RA, http://scs.am/am/6954e433a4402db729623210

In the first stage quantitative measurements have been carried out to find out the total number of signed international bilateral documents in science, their dynamics, regional orientation (grouping the states by regions and unions) the specific weight of each group in the total. Using the method of cluster analysis, we divided the states into groups and conducted a comparative analysis.
In the second stage, competitions jointly held by Armenia and other states have been analyzed. To determine the share of specialties in the total number, the full count method was used. (Robertson et al., 1980) In other words, one point was given to each specialty for the opportunity to participate in the competition. Then, adding up the points received for the opportunity to participate in all competitions, and comparing with the total number of points, the percentage weight of each profession in total was obtained. Thus, the priorities of professions have been determined. Then, to determine the classification of specialties, the total scores of all specialties in a given area have been correlated with the number of specialties. After that, the indicators obtained at this stage have been compared with those retrieved at the first stage in order to analyze the applicability and viability of the signed international instruments.
In the third stage the results of the competitions have been analyzed subjecting the winning projects to quantitative measurements. By applying the method presented for the second stage, it has been revealed which scientific fields have more selected projects, and how priorities have changed due to the regional cooperation. The results obtained in this round have been compared with the results of the previous two rounds.
In the fourth stage we have discovered which projects have ended up with publications in the journals indexed in Web of Science. It helped to find out the correlation between the winning projects and the number of articles published in well-known international journals, the overall dynamics of articles' publication, the fields of science, the number of received citations.
All specialties are grouped into 6 major scientific fields and 36 subfields using the Frascati classification. There are slight changes from the original version of Frascati classification, presented for the first time in 1963.[7] Originally there are 42 scientific subfields (Kutlača et al., 2015) and Science Committee uses 36 out of them. Also it has included Armenian Studies or Armenology it this classification. The six main scientific fields are: 1) Natural Sciences (mathematics, informatics and computer science, physics and astronomy, chemistry, geosciences and related environmental sciences, biological sciences), 2) Engineering and Technology (urban planning and architecture, computer science and information technology, mechanics, machine science and mechanical engineering, chemical technology, materials science, medical instrumentation, ecology, biotechnology, nanotechnology), 3) Medical Sciences (general medicine, clinical medicine, medical biotechnology), 4) Agricultural Sciences (animal husbandry and veterinary medicine, horticulture, soil science and plant protection, agricultural biotechnology), 5) Social Sciences (psychology, economics and business, pedagogical sciences, sociology, law,

[7] Detailed information can be found in Frascati Manual 2015.

political sciences, social and economic geography, media), 6) Armenology and the Humanities (history and archeology, linguistics and literary criticism, philosophy and ethics, theology and religious studies, art history).

Results and discussion

After declaring independence, the Republic of Armenia, as an independent subject of international relations, is active in forming a new network of international cooperation. 102 bilateral and multilateral agreements with foreign partners and international scientific organizations were signed by the Government of RA and the authorized state body responsible for science between the years 1991 and 2022 aimed at establishing cooperation in the field of science, upgrading the local science to the international standards and integrating it into the international scientific community (*Appendix 1*). This table reveals that the process of establishment of interstate scientific ties was permanent, continuous and expanding in nature. Only in 2004-2007 the process has stalled and no contracts have been signed. It is difficult to give an exact explanation what caused this, but it could be claimed that the implementation of preparations for the transition to a qualitatively new phase played an important role in it.

Figure 1 revealed that in the field of scientific collaboration RA has had multi-vector orientation. In consequence of simple comparison of the number of states the first place belongs to the EU countries, followed by CIS and Asian ones. But it should be mentioned that EU's first place is due to the fact that it has more member states than CIS. Moreover, Armenia has bilateral and multilateral agreements with almost all CIS member states,[8] whereas in the case of EU this rate is 60%. And in the case of Asian countries the percentage is much lower.

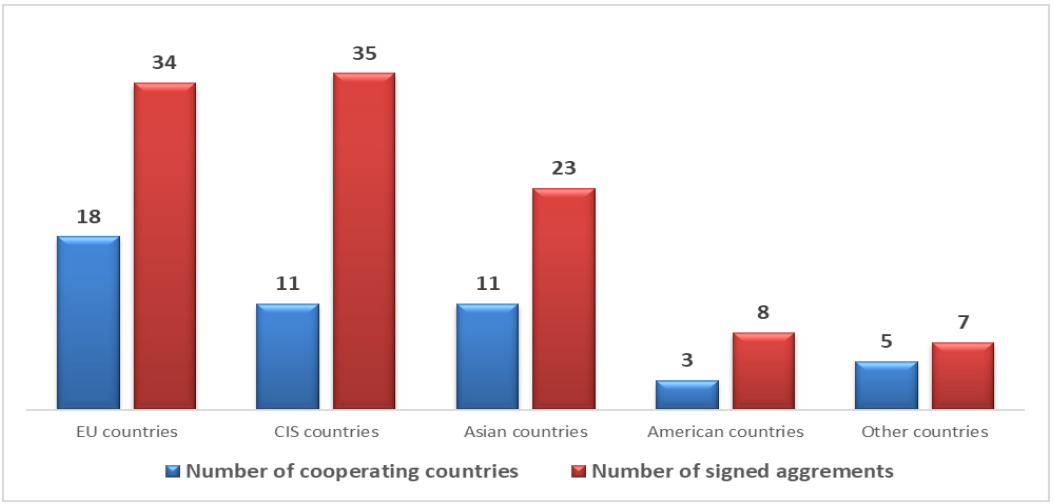


Figure 1. International agreements of RA in the field of science.

[8] There are no bilateral agreements only with Azerbaijan and Moldova. The reason for it is the lack of diplomatic and good neighborly relations in the case of the former, and the passivity and lack of interest in the bilateral relations in the case of the latter.

Figure 1 also shows that the closest, deepest and most multi-vector relations have been established with CIS countries. This is due to the longstanding historical, cultural, political, regional and lingual cooperation and mutual relations between Armenia and CIS countries. On the other hand there are continuously developing relations with EU countries. There is also a trend of deepening the relations with developed and developing Asian countries. And when it comes to the interstate scientific relations with American countries it should be mentioned that they are developing very slowly.

At the level of bilateral relations, the largest number of agreements was signed with Russia – 15, followed by the Republic of Belarus with 6 agreements when considering the CIS countries. Among the EU countries, relations with Italy (8 agreements), Romania (4 agreements), France (2 agreements) and Germany (2 agreements) are more active. The number of agreements signed with these countries testifies the deepening and intensifying nature of scientific ties with them, since these agreements include and regulate various aspects of cooperation.

It should be noted that not all signed interstate agreements were implemented, and some of them had no results. In order to eliminate this negative phenomenon, as well as to increase the coordination of science and regulate the state support provided to science, the Science Committee was created under the Ministry of Education and Science in 2007. Thanks to its efforts, international agreements on scientific cooperation lead to significant results. The first result is bilateral competitions between Armenia and other foreign countries, through which numerous of scientific projects have been financed.

During its activity the Science Committee signed 42 international agreements in the sphere of scientific cooperation, on the basis of which 35 competitions were realized. This is about 36% of the total number of competitions[9] organized by the Science Committee. According to the results of international competitions the winning projects received short-term funding, mainly for 12 or 24 months.

Figure 2 shows that international bilateral scientific competitions were held mainly with CIS and EU countries, that is why in the Figures 3 and 4 we have concentrated only on these two groups of countries. In other words, the above-mentioned arrangements with Asian, American (except Canada) and other countries did not lead to the announcement and holding of joint competitions. In total 25 competitions were organized with CIS countries and 8 with EU countries (one Armenian-German and one Armenian-Italian competitions have been summarized recently but they are out of the time span of the article and due to it have not been considered). Both in terms of the signing of interstate agreements and the organization of joint competitions, among the CIS countries the most active relations are with Russia and Belarus, and among the EU countries - with France, Germany and Italy. Figure 2 also shows that 2016 was the most productive year in terms of international competitions: 8 competitions were organized with both CIS and EU countries, as well as with Canada.

[9] 93 domestic and international competitions were held by the Science Committee in RA in order to finance scientific programs.

The next fact that becomes clear from Figure 2 is that the established bilateral relations are mostly developing and continuous. The bilateral Armenian-Russian competitions, which began in 2012, continue to this day: 10 calls have been announced since then. The same picture is with the Armenian-Belarusian competitions, which started in 2011 and were held 7 times. France was the first EU country to organize bilateral competitions, but this process was interrupted in 2013. Joint Armenian-German and Armenian-Italian competitions have continuous nature. But they are organized with long interruptions.

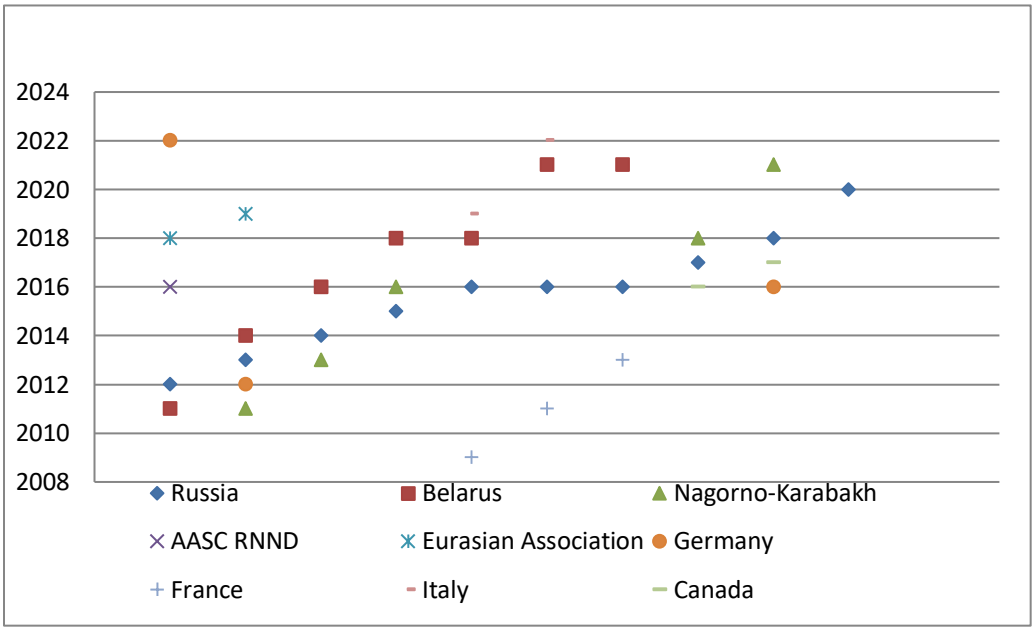


Figure 2. Timetable of international scientific competitions.

The announced international competitions differ in purpose, focus, and preferred specialties. Accordingly, in terms of eligibility, projects in different specialties have unequal opportunities for participation. Moreover, the picture is different when looking at the CIS and EU countries separately. Figure 3 shows that the biggest number of international competitions belongs to natural sciences, 21 of which were for biological sciences (16 competitions with CIS countries and 5 with EU countries). Next are Physics and Astronomy - a total of 20 competitions (15 times with CIS countries and 5 times with EU ones). Chemistry and Earth Science had a little bit less opportunities. Agricultural sciences had the least chance to participate. Projects in Mechanics, Mechanical Engineering, Chemical Engineering, and Ecology had not opportunities to participate in competitions with EU countries. History and Archaeology received the most opportunities among the Humanities - 16 competitions (9 times with the CIS countries and 7 times with the EU countries), and among the Social Sciences Sociology is the leader - 15 competitions (9 times with the CIS countries and 6 times with the EU countries). This picture shows the degree of development of science in Armenia and the range of interests with foreign countries.

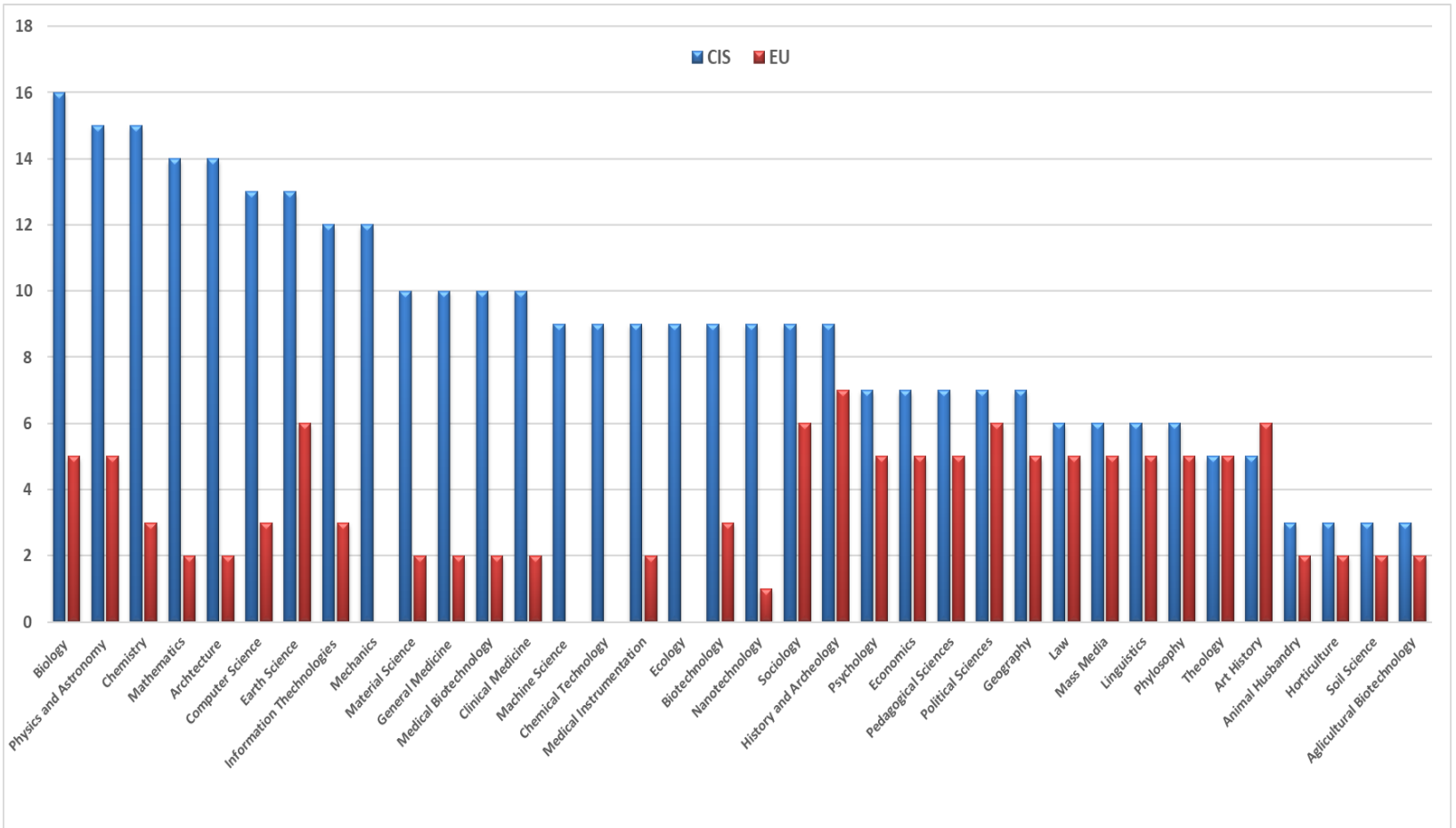


Figure 3. International scientific competitions by specialties.

If we compare the total number of opportunities to participate with the total number of international competitions, we see that in the case of Biological Sciences the opportunity to participate is 60%, Physics and Astronomy is 57.5%, Earth Science – 54.2%, Chemistry - 51.4%, History and Archaeology - 45.7%, Sociology - 42.8%, and in the case of Agricultural Sciences - 14.2%.

A total of 1,089 projects have been submitted to the 33 international competitions organized and held so far, of which 332 (30.4%) won and received funding. 119 of these programs have been submitted to bilateral competitions with EU countries, of which 47 (39.4%), have been announced as winners. Of the 871 projects submitted to bilateral competitions with CIS countries 242 (27.7%) were winners. The most effective were the bilateral competitions organized by the SCS RA and RFBR RF, in which 556 projects participated, of which 119 (or 21.4%) were guaranteed for funding.

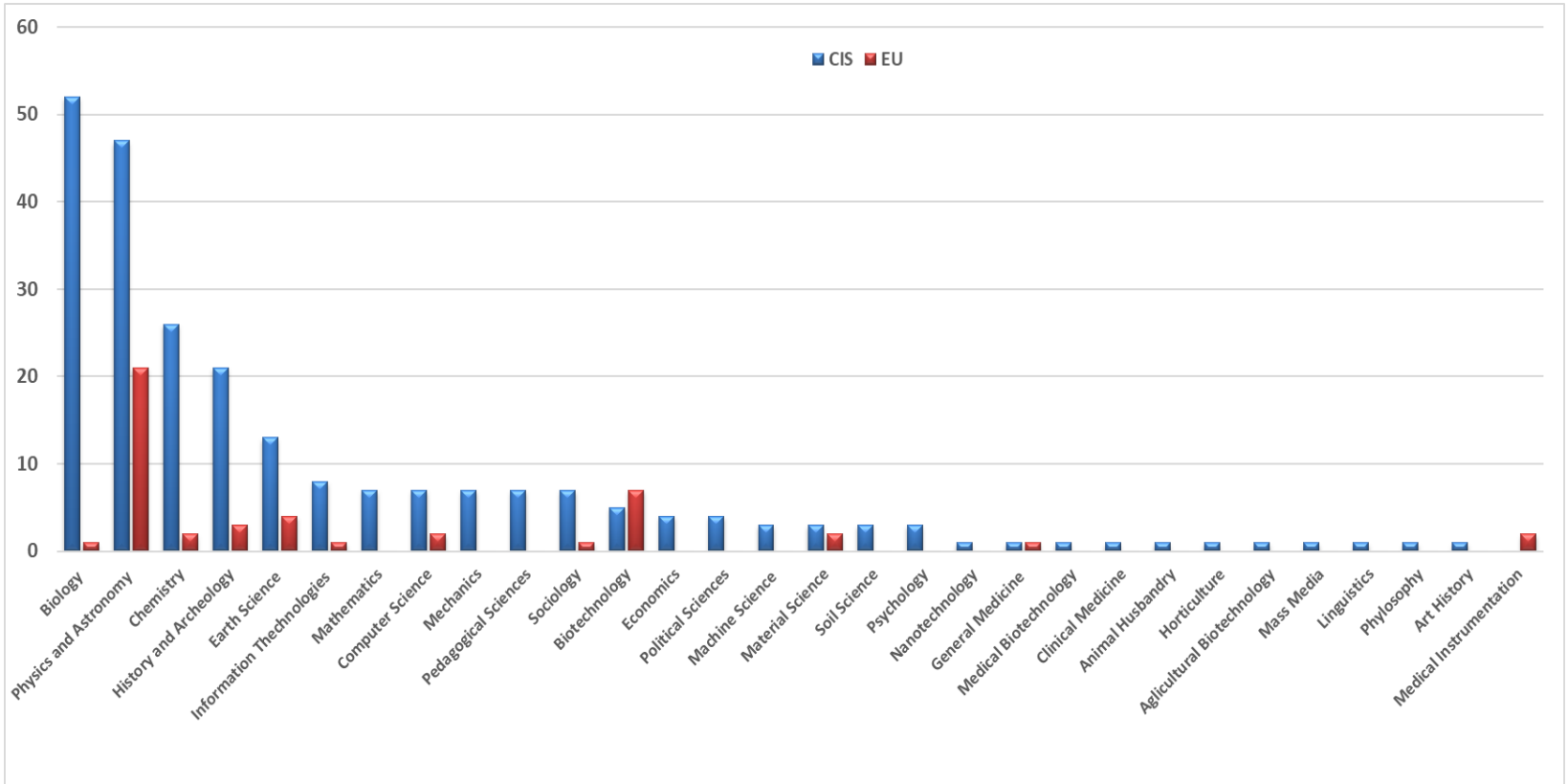


Figure 4. The total number of winning projects in international competitions by specialties.

Figure 4 shows that in bilateral competitions organized with CIS and EU countries, the greatest number of winning projects belongs to natural sciences - 213 ones which is 64.1% of all financed projects. Biology ranks first in the number of winning projects in competitions held with CIS countries, with 52 projects receiving funding. It is followed by Physics and Astronomy and Chemistry with 47 and 26 winning projects respectively. History and Archeology are leaders (21 projects) in the field of Humanities in the terms of the number of winning projects. Among the winning projects in joint competitions held with the EU countries the absolute leaders are Physics and Astronomy (21), followed by Biotechnology (7). For a number of specialties there were no winning projects at all (Ecology, Law, Geography, Theology, Architecture, Chemical Technology).

One of the important achievements of the international competitions is thousands of articles written in co-authorship by Armenian and foreign scientists. Most of them have been published in journals indexed in international scientific databases. For example, in the frames of 106 winning projects 377 publications have been published in journals indexed in the Web of Science database, of which 351 are articles. 341 of these publications are the result of collaborative work. It should also be mentioned that the collaborations established through the competitions have influenced the publication of joint new articles beyond these competitions. This fact is the further evidence that the number of joint publications by Armenian scientists with their foreign colleagues has been growing steadily in recent years.

Figure 5 reveals that the number of articles is increasing significantly. At the same time, the number of articles published in recent years has increased several times compared to the first years. All articles were published in English, except one, which published in Russian.

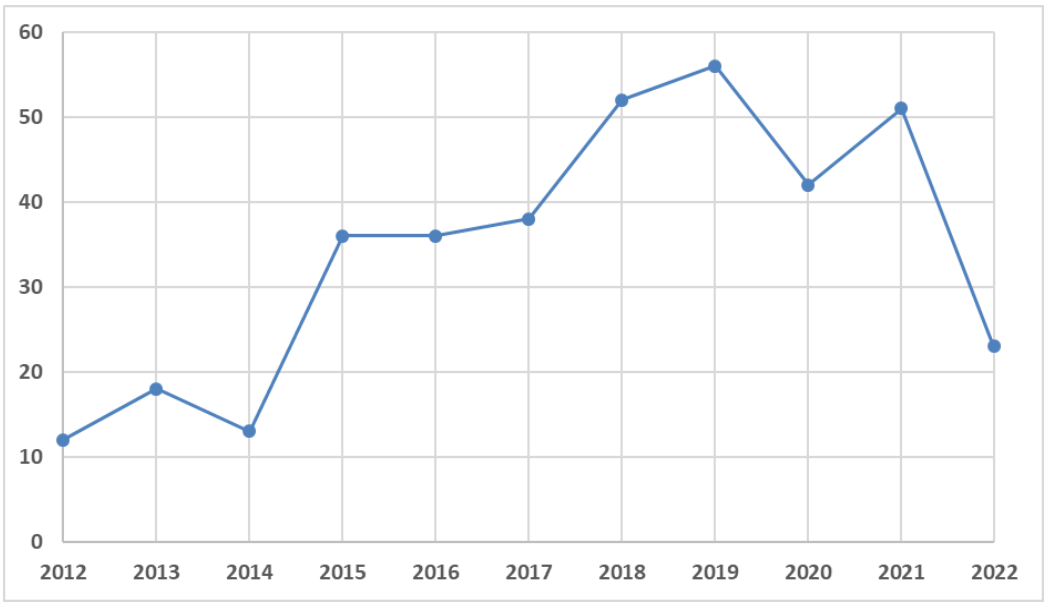


Figure 5. The articles published in the frames of bilateral competitions in the WOS indexed journals.

Figure 6 shows that Armenian scientists collaborate mainly with their Russian colleagues. Armenian and Russian scientists are co-authors of 46.4% of the articles published in WOS in the frames of the winning programs of bilateral competitions. This is logical, since Russian-Armenian bilateral competitions and winning-projects have a large share in the total volume. This collaboration is followed by cooperation with Germany, Belarus, France and Italy. The noteworthy fact is that although bilateral competitions were held more often with Belarus than with Germany, and more projects were guaranteed for funding, cooperation with German scientists is more intensive.

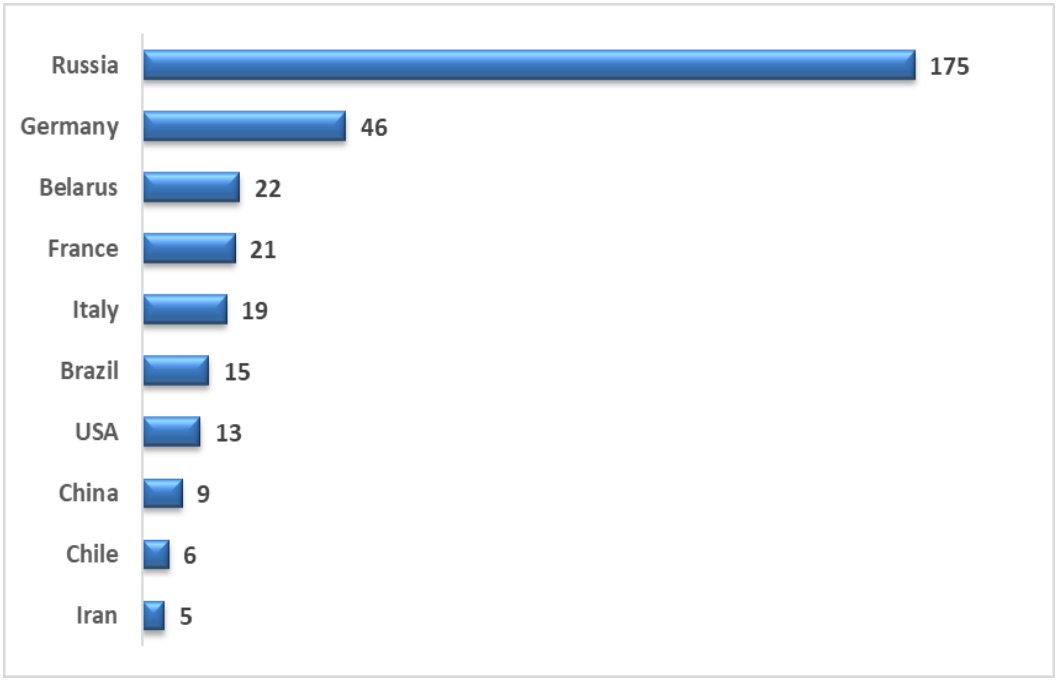


Figure 6. Top 10 countries of scientific collaboration based on joint publications of winning projects.

Table 3 shows that international cooperation is especially active in the field of Physics and its results exceed the total result in other areas. Articles published in this area of research account for 53.8% of the total number of articles. Physics is followed by Chemistry (10.3% of published papers), then by material science (9.8%) and mathematics (9.2%). An interesting picture emerges when comparing the number of published articles with the number of winning projects (Fig. 4). Physics is in first place both in terms of the number of winning projects and the number of published articles. There is a tiny gap between physics and biological sciences in terms of winning-projects number. Although a significant amount of articles have been published in the field of biological sciences it cannot be compared with the number of published articles in the field of physics. There is a controversial picture when comparing material science with history and archaeology. In the case of former, 37 papers were published in the frames of 5 winning projects. And in the case of latter

no articles have been published in any of WOS indexed journals although this field ranks 4th in the terms winning projects number - 24.

One of the most important ways to measure the quality characteristics, importance, and applicability of articles is the number of received by citations. From 377 published articles only 270 received citations. Table 1 shows that the total number of citations received by articles published in the field of Physics is extremely high – 1491. Although the total number of articles published in the field of Chemistry was second only to physics, but in terms of the number of citations received (221), it is also inferior to Material Science (410) and Science Technology Other Topics (300). Articles published in the field of Chemistry are in the middle positions in terms of the average value of the received citations. According to this indicator the leaders are Material Science (11.08), Engineering (11.36), and Environmental Sciences Ecology (13.3).

Table 1. Number of publications and received citations by Web of Science Subject Categories.

WOS subject categories	Number of Publications	Number of Received Citations
Physics	203	1491
Chemistry	39	221
Material Science	37	410
Mathematics	35	98
Science Technology Other Topics	32	300
Optics	31	185
Astronomy Astrophysics	26	107
Biochemistry Molecular Biology	13	66
Engineering	11	125
Environmental Sciences Ecology	10	133
Geology	8	37
Pharmacology Pharmacy	7	35
Computer Science	6	15
Genetics Heredity	6	45
Geochemistry Geophysics	6	51
Radiology Nuclear Medicine Medical Imaging	6	9
Biothechnology Applied Microbiology	5	38
Biophysics	4	24
Crystallography	4	21
Life Sciences Biomedicine Other Topics	4	2

If we consider the number of articles published in the frames of the winning projects and the number of received citations by group of countries (Figure 7) it will become apparent that the cooperation with CIS countries is in first place with 190 published articles and 990 received citations. This is followed by cooperation with the EU countries with 101 articles and 904 received citations. It should be noted that the citation per publication is greater for EU countries than CIS. It is interesting that bilateral competitions were not held with the countries of America and Asia, but articles were published in co-authorship with scientists from those countries. At the same time, publications in co-authorship with scientists from American countries are in first place in terms of the average index of citations received by them.

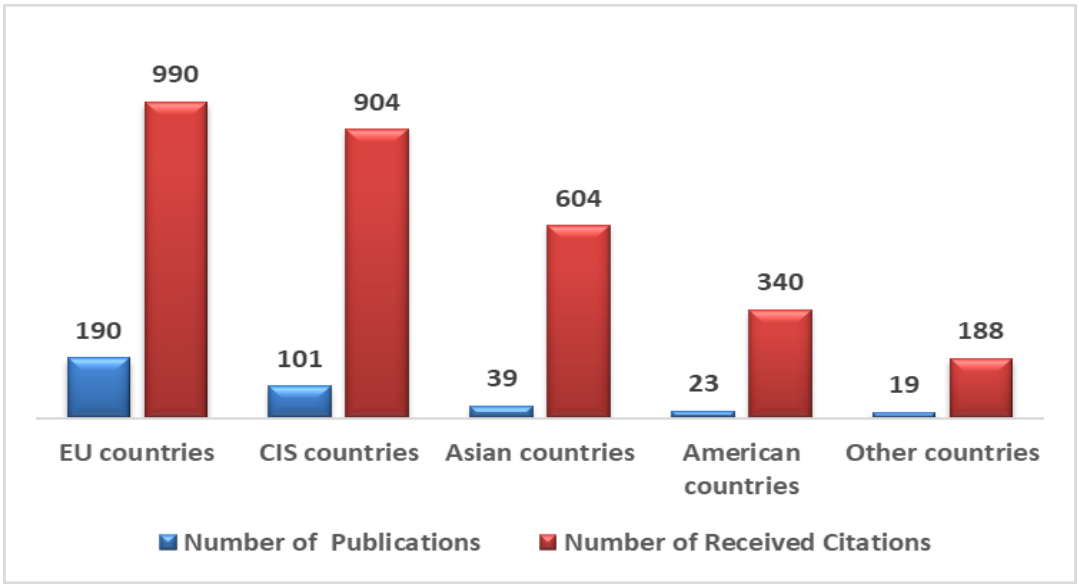


Figure 7. The number of publications and citations by groups of countries.

Conclusion

International scientific cooperation can be measured using different methods (Wang L. et al., 2017). Analyzing international scientific cooperation in Armenia from the point of view of state-established international scientific cooperation and the results achieved due to it, we found out that the general state policy in Armenia was aimed at internationalizing science, localizing international scientific achievements and experience, using new methods and means, and integrating into the international scientific market.

After gaining independence, the Republic of Armenia pursued a multi-vector policy of establishing scientific ties. On the one hand the signed agreements preserved and developed relations with the former Soviet states. In particular, bilateral scientific cooperation has been established with the Russian Federation and the Republic of Belarus. On the other hand, taking advantage of the opportunity to pursue an independent policy, new scientific ties were established with dozens of countries in Europe, America and Asia. Thus, local science received a new impetus, overcame a

number of outdated frameworks. Furthermore, it was created an opportunity to discuss many issues from the fundamental points of view which are common in modern world.

2009 was an important milestone in the further development of international scientific cooperation, when international bilateral scientific competitions were launched. They provided an opportunity for Armenian and foreign scientists to form research groups and work together on the implementation of various scientific projects. The number of such competitions has increased over the years, leading to more interest in them. During the implementation of projects, Armenian scientists had the opportunity to cooperate with their colleagues from France, Germany, Italy, Russia and Belarus and learn from their experience. Joint efforts were directed to new research and discoveries. Through collaboration, partners can share knowledge, skills, and techniques and improve productivity (Katz & Martin, 1997). In general, there was an interaction of scientific cultures with all positive consequences.

In the course of this study, it became clear that international scientific collaboration is developing most actively in the field of natural sciences and they have more opportunities both in terms of participation in competitions and in terms of success in them. It is no coincidence that the lion's share of the winning-projects falls on ones in the fields of biology, physics and astronomy, chemistry, and earth sciences. Among the humanities, there is a strong interest in the field of history and archaeology. In the frame of bilateral competitions with the CIS countries it had a significant success.

Hundreds of co-authored articles published in journals included in international scientific databases are also among the important outcomes of international bilateral competitions. For example, 377 papers published in the frames of bilateral competitions can be found in Web of Science, 341 of which are co-authored. More than half of the articles are in physics.

References

Adams, J. (2013). The fourth age of research. *Nature,* 497, 557-560. https://doi.org/10.1038/497557a

Astakhova, M. (2020). Scientific Cooperation Across the BRICS. *BRICS Law Journal,* 7(1), 4–26. https://doi.org/10.21684/2412-2343-2020-7-1- 4-26

Bote, V.P.G., Olmeda-Gómez, C., de Moya-Anegón, F. (2012). Quantifying the benefits of international scientific collaboration. *Journal of the American Society for Information Science and Technology*, 64(2), 392-404. https://doi.org/10.1002/asi.22754

Coccia, M., Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *PNAS*, 113(8), 2057-2061. https://doi.org/10.1073/pnas.1510820113

Finardi, U., Buratti, A. (1999). Scientific collaboration framework of BRICS countries: an analysis of international coauthorship. *Scientometrics,* 109(1), pp. 433-446. https://doi.org/10.1007/s11192-016-1927-0

Freeman, R. (2010). Globalization of scientific and engineering talent: international mobility of students, workers, and ideas and the world economy. *Economics of Innovation and New Technology,* 19(5), 393-406. https://doi.org/10.1080/10438590903432871

Gazni, A., Didegah, F. (2011). Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. *Scientometrics*, 87(2), 251-265. https://doi.org/10.1007/s11192-011-0343-8

Gazni, A., Sugimoto, C.R., Didegah, F. (2012). Mapping World Scientific Collaboration: Authors, Institutions, and Countries. *Journal of the American Society for Information Science and Technology*, 63(2), 323–335. https://doi.org/10.1002/asi.21688

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69–115. https://doi.org/10.1023/A:1010512628145

Glukhov V.A., Gzoyan E.G., Sargsyan Sh.A. Assessment of Scientific Cooperation between the Scientists from Armenia and Russia within the Joint Bilateral Grant Projects. *Sociological Studies*, 7, 156-158. (In Russ.) https://doi.org/10.7868/S0132162517070182

Gomez, I., Teresa Fernandez, M., Sebastian, J. (1999). Analysis of the structure of international Scientific cooperation networks through Bibliometric indicators. *Scientometrics*, 44(3), 441-457. https://doi.org/10.1007/BF02458489

Gui, Q., Liu, C., Du, D. (2019). Globalization of science and international scientific collaboration: A network perspective. *Geoforum*, 105, 1-12. https://doi.org/10.1016/j.geoforum.2019.06.017

Gzoyan E.G., Mirzoyan A.R., Aleksanyan S.A., Oganesyan L.A., Unanyan S.R., Megrabyan M.M., Glukhov V.A., Sargsyan S.A. (2017) The role of state grants in the Armenian-Russian scientific ties development: bibliometric analysis. *Bibliosphere*, 3, 69-77. (In Russ.) https://doi.org/10.20913/1815-3186-2017-3-69-77

Katz, J. S., Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1-18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kutlača, D., Babić, D., Živković, L., Štrbac, D. (2015). Analysis of quantitative and qualitative indicators of SEE countries scientific output. *Scientometrics*, 102(1), 247–265. https://doi.org/10.1007/s11192-014-1290-y

Moed, H.F., de Bruin, R.E., Nederhof, A.J. Tijssen, R.J.W. (1991). International scientific co-operation and awareness within the European community: problems and perspectives. *Scientometrics*, 21(3), 291-311. https://doi.org/10.1007/BF02093972

Nagpaul, P.S. (2003). Exploring a pseudo-regression model of transnational cooperation in science. *Scientometrics*, 56(3), 403–416. https://doi.org/10.1023/A:1022335021834

OECD (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, The Measurement of Scientific, Technological and Innovation Activities. OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264239012-en

Pfotenhauer, S.M., Wood, D., Roos, D., Newman, D. (2016). Architecting complex international science, technology and innovation partnerships (CISTIPs): A study of four global MIT collaborations. *Technological Forecasting & Social Change*, 104, 38-56. http://dx.doi.org/10.1016/j.techfore.2015.12.006

Robertson, S. E., van Rijsbergen, C.J., Porter, M.F. (1980). *Probabilistic models of indexing and searching*, SIGIR 80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval, pp. 35–56.

Ruffini, P.-B. (2017). *Science and Diplomacy: A New Dimension of International Relations (Science, Technology and Innovation Studies)*. Springer.

Sargsyan, Sh.A., Maisano, D.A., Mirzoyan, A.R., Manukyan, A.A., Gzoyan, E.G. (2020). EU-EAEU dilemma of Armenia: Does science support politics? *Scientometrics*, 122, 1491–1507.

Schott, T. (1991). The world scientific community: Globality and globalization. *Minerva*, 29, 440–462. https://doi.org/10.1007/BF01113491

Schmoch, U., Schubert, T. (2007). Are international co-publications an indicator for quality of scientific research? *Scientometrics*, 74(3), 361–377. https://doi.org/10.1007/s11192-007-1818-5

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. *Scientometrics*, 81(1), 177–193. https://doi.org/10.1007/s11192-009-2126-z

Shugurov, M.V. (2019). On the Issue of Supranational Aspect of Legal Regulation of Innovative and Scientific-technological Cooperation of the EAEU Member States. *Russian-Asian Law Journal*, 2, 74-79. (In Russ.)

The Royal Society. (2011). *Knowledge, networks and nations. Global scientific collaboration in the 21st century.* RS Policy document 03/11.

Wang, L., Wang, X., Philipsen, N.J. (2017). Network structure of scientific collaborations between China and the EU member states. *Scientometrics*, 113(2), 765–781. https://doi.org/10.1007/s11192-017-2488-6

Wang, Y., Long, Y., Tu, L., Liu, L. (2022). Delivering Scientific Influence Analysis as a Service on Research Grants Repository. *IEEE Transactions on Services Computing*, 15, 1896-1911. https://doi.ieeecomputersociety.org/10.1109/TSC.2020.3025318

Wardil, L., Hauert, Ch. (2015). Cooperation and coauthorship in scientific publishing. *Physical Review E*, 91, 012825. http://dx.doi.org/10.1103/PhysRevE.91.012825

Appendix

International agreements signed by Republic of Armenia with other countries and organizations.

1991	Romania	2010	Czech Republic
1992	China		Kazakhstan
	Vietnam		Belarus
	Iran		Russia
	Argentina		Russia
1993	Russia		Belarus
	Russia	2011	CIS
	Turkmenistan		Russia
	Georgia		Russia
1994	Argentina		Germany
	Romania		Italy
	Greece		Germany
	United Kingdom		Switzerland
1995	Romania	2012	Lithuania
	Iran		Russia
	India		China
	Israel		China
	Russia		Turkmenistan
	Lebanon		Vietnam
	INTAS international organization		Spain
	Russia		Iraq
1996	France	2013	Russia
	Ukraine		Italy
1997	USA		Italy
	Egypt		Belarus
	CIS		Belarus
	Kyrgyzstan		Ukraine
1998	Georgia	2014	Estonia
	Ukraine	2015	China
	Poland		Italy
	International Science and Technology Center		EU, Georgia, Japan, Kingdom of Norway, Kyrgyzstan, Kazakhstan, Korea, Tajikistan, USA
	Cyprus	2016	Russia

	Lebanon		India
	Georgia		Russia
2000	Slovakia		Russia
	Portugal		Georgia, Moldova, Ukraine, Azerbaijan
	Belarus		Belarus, Vietnam, Mongolia, Kyrgyzstan, Russia
2001	Russia		Canada
	Romania		Korea
	Bulgaria		EU
2002	India	2017	Bulgaria
2003	Tajikistan	2018	Italy
	Italy	2019	Vietnam
2008	Russia		Italy
2009	Croatia		China
	USA		Canada
	Kuwait		CIS
	Latvia		Russia
	France	2020	Canada
	International Science and Technology Center	2022	China
2010	Slovenia		Italy

https://doi.org/10.51408/issi2025_085

Scientific Travelers Associated with Less Disruption but Better Scientific Novelty

Mingze ZHANG[1], Penghui LYU[2], Yizhan LI[3], Zexia LI[4]

[1] *zhangmingze@mail.las.ac.cn,* [3] *liyz@mail.las.ac.cn* [4]*, lizexia@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[2] *sibiling@uestc.edu.cn*
Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, 518000 (P.R. China)

Abstract

Building on the framework of facilitymetrics and the features of big science facilities, this study provides a more micro method to identify the scientific mobility procedure, named scientific travels hereafter, and associated with scientific performance at the author level and paper level. We classify external users of big science facilities into two types (travelers and locals) by measuring the number of facilities the focal scientist's used, measured by co-authored publications, during a specific period (one year, previously, and career level), visualize their gap in scientific performance, which is measured by a five-year disruption index and novelty score, and validate the impact relationships by causal inference respectively in paper-level and author-level. Results show that locals might produce more disruptive knowledge while travelers perform better in novel knowledge production. Paper-level and author-level regressions validate the results that the participation of travelers in teams leads to better novelty but lower disruption, and the performance gaps between travelers and locals surely exist. However, from a long-term perspective, the disruptive ability could increase significantly as a traveler is fully localized and gradually surpasses his or her peers' ability. The novelty ability of travelers might decrease slowly but insignificantly since they are always ahead of locals and their peers. This study contributes to understanding the performance evaluation and science policy in big science facilities, which enriches the research in scientific mobility, and the results could be a reference for those short periods of scientific activity related to mobility without visible information to map and quantify.

Introduction

Scientific mobility is highly motivated by the development of transportation and the trends of globalization (Lin, Frey, & Wu, 2023), especially since the 21st century. Scientists, with their knowledge, can travel around the globe easily, communicate with distant peers, collaborate for new progress, and chase career success (Wang, Hooi, Li, & Chou, 2019). High mobility has already transformed the paradigms of knowledge production by several approaches, for instance, local knowledge could flow to a wider academia easily, and knowledge from different regions could be highly connected for global scientific progress (Franzoni, Scellato, & Stephan, 2012; Söderström, 2023a). As for a scientist, he or she could serve as a carrier of regional knowledge outflows to global academia. Similarly, scientists could be trained in multi-regions and eventually bring his or her diverse knowledge to in-flow regions (Thelwall & Maflahi, 2022).

In the science of science, the performance of scientific mobility receives great attention, and many studies are demonstrating the benefits of scientific mobility (Aykac, 2021; De Filippo, Casado, & Gómez, 2009). Even though temporary performance loss at individual and collective levels (so-called brain drain) is reported (Abramo, D'Angelo, & Di Costa, 2022; Verginer & Riccaboni, 2021) and types of inequality exist concurrently (Deville et al., 2014; Gu, Pan, Zhang, & Chen, 2024; Momeni, Karimi, Mayr, Peters, & Dietze, 2022), scientific mobility is still considered an effective way to improve individual performance in impact and productivity and is beneficial to returnees' regions for a long-term perspective (Holding, Acciai, Schneider, & Nielsen, 2024; Liu & Hu, 2022).

Thus, we suppose that the identifications of scientific mobility are not able to keep up with the increasingly evaluating demands in short-term scientific travels for communication and collaboration. Concurrently, most identifications based on the changing information in individuals' affiliations and the related data are always extracted from their published records, scholar identity, and self-disclosing Curriculum vitae (CV). Such methods are still at a coarse-grained level since they might neglect several short-term scientific movements, which might also influence individual performance. We collected a unique dataset from the publications of global big science facilities, which could be used to fill this knowledge gap.

Big science is considered one of the basic features of modern science, and big science facilities are research infrastructures for modern science. National or supranational bodies began the investments during World War II and are expecting these big machines to assist cutting-edge knowledge discoveries with advanced analytical technologies, especially in science-related disciplines (Hallonsten, 2014). Nowadays, big science facilities are operated as user-oriented experimental platforms, which requires users, considered as external scientists, from global academia to submit their research proposals and conduct their experiments on-site if users' proposals are permitted successfully (Heinze & Hallonsten, 2017; Silva, Schulz, & Noyons, 2019; Söderström, 2023b).

The utilization model of big science facilities provides us with a novel perspective to identify scientific mobility in a more micro way, and we suppose that "scientific travel" is a more suitable concept (Söderström, 2023a). Therefore, we demonstrate that those co-authored external scientists of the facility could be defined as scientific travelers if they are recorded in more than one facility during a specific period, and they are considered as scientific locals if they are only recorded in one facility. After the classification, we could compare the performance gaps between two types of external users at the author level and paper level by measuring the disruptive and novel abilities.

This study contributes to current knowledge in several ways. Firstly, we proposed a more micro way to identify scientific mobility and named such level movements as scientific travelers, enriching the current research on the relationships between scientific performance and scientific mobility from a novel and unique perspective based on the research context in big science facilities. Secondly, we contribute to expanding the framework of facilitymetrics by providing significant evidence related to the performance gaps between different types of users (diverse or concentrated)

to facilitate the practices of brain gain and science policy in the era of big science. Thirdly, the results from the micro perspective could be extrapolated to those short-term scientific activities full of knowledge communication and peer collaborations but concurrently hard to be identify in the level of scientific big data, for instance, attending international conferences, the plans of visiting scholars, and other on-site collaborations with cross-regional co-authors.

We review the extant literature related to big science facilities and scientific mobility, introduce our methods of data collection, indicator construction, and quantitative predisposition, display our main results and supporting results, and discuss the potential implications of our results to science policy in the following sections.

Literatures Review

Big Science Facilities and Facilitymetrics

Big science is a concept that has already existed for at least several decades since World War II, which gave birth to a group of research infrastructures with advanced experimental technologies and unique scientific circumstances for cutting-edge knowledge discoveries in science disciplines (Hallonsten, 2016; Heinze & Hallonsten, 2017). Such research infrastructures, named big science facilities, are commonly invested by national or supranational bodies since the processes of construction and maintenance require too much vast investment, huge network resources, and collective efforts to be afforded by one or several universities and institutions (D'Ippolito & Rüling, 2019). Therefore, the nature of big science facilities contain the concept of shared and are ready to open for scientific progress (Hallonsten & Christensson, 2017; Lauto & Valentin, 2013), known as user-oriented, and should be responsible to their taxpayers since they are public investment goods. Under such context, one cutting-edge discipline, so-called facilitymetrics, arose and has already developed for a decade to apply, revise, and update quantitative methods from scientometrics to evaluate the scientific performance of big science facilities. Facilitymetrics is first proposed by Hallonsten (2013) with suitable indicators (Hallonsten, 2014), for instance, Facility Immediate Index and Facility Impact Factors (Heidler & Hallonsten, 2015), applied to evaluate these machines' performance based on the scientific publications supported by them.

The development of facilitymetrics originated from the special features of big science facilities, leading to the evaluations of scientific performance should consider those hidden factors. For instance, the extreme number gap between investments and productivity might lead to absurd evaluative results (Lauto & Valentin, 2013). Moreover, in the context of big science facilities, knowledge production is highly depended on collaborations and the collaboration between communities should be highlighted since there are two unique communities of scientists related to big science facilities, named external scientists (users) and internal scientists (staff), respectively. With respect to previous studies in theories and the user orientation in practices, such a collaboration paradigm might damage the research chances of internal scientists and emphasize their functions of supporting and serving, which placed them in an underrepresented condition

(D'Ippolito & Rüling, 2019; Silva et al., 2019; Söderström, 2023b). However, in our previous work, results demonstrated that the paper-level performance would be significantly improved if external users collaborate with those internal scientists, ensuring the indispensable effects of internal scientists. From the theories of team science (Katz & Martin, 1997; van Knippenberg & Schippers, 2007), we supposed that it might be the heterogeneous knowledge, for instance, technology manipulation or data interpretation, that internal scientists possess that makes collaborative users conduct their experiments easier, more effective, and more standardized (Xu et al., 2024; Yang, Tian, Woodruff, Jones, & Uzzi, 2022). Eventually, succeed in scientific performance.

The utilization of most big science facilities is on-site (Söderström, 2023a), but these facilities are still suffering from the shortages of beamtime and research resources since the booming demands from global users and the annual experimental volumes in one facility are limited by natural reasons (D'Ippolito & Rüling, 2019). Therefore, potential users are required to submit research proposals to compete and await to be permitted by facilities (Hallonsten & Christensson, 2017). Those successful users need a short period to visit the facility and finish their research on-site during the limited beamtime. Such a mechanism enables us to identify whether the focal author traveled or not during a specific period.

After all, big science facilities are considered experimental platforms for scientific research, especially important for those disciplines that highly depend on advanced analytical technologies such as X-rays, Particle accelerators, Free-electron lasers, and Neutrino detectors. Therefore, there are different types of big science facilities, and Synchrotron Radiation Lightsource (SLS) is one of the most attractive facility types in the framework of facilitymetrics. It is reported that about 50 SLSs are operating, and some of them are still under construction around the world concurrently (Conroy, 2024; Wild, 2021), and most of them have already produced considerable scientific knowledge with several Nobel prizes related to (Hand, 2010; Heinze & Hallonsten, 2017; Jiménez, 2010). Therefore, we mainly focus on the performance of SLSs in this study and confine our focal scientists to the community of external scientists for high accuracy to define travelers and locals with respect to the unique features abovementioned in the context of a big science facility.

Scientific Mobility and Individuals' Performance

One of the features of modern science that benefited from the development of transportation is that scientific individuals could move around the globe more easily than before to communicate and collaborate with their peers (Franzoni et al., 2012; Lin et al., 2023; Söderström, 2023a; Van Noorden, 2012). Many studies have provided evidence to demonstrate the impacts of scientific mobility, and such influences could be divided into two aspects approximately. One focuses on evaluating the socio-economic impacts and the future of in-flow and out-flow regions (Verginer & Riccaboni, 2021) and the other attempts to discover the variation of individuals' scientific performance (De Filippo et al., 2009).

In the science of science, scientific mobility is tightly associated with the evaluations of scientific performance (De Filippo et al., 2009). Moving to another place might

bring several risks and challenges (Deville et al., 2014), leading to a temporary productivity loss (Abramo et al., 2022), disconnecting with previous colleagues in the former affiliations gradually (Wang et al., 2019), and eventually damaging individuals' performance. However, from a further perspective, specifically at the career level, the main viewpoint of scientific mobility demonstrates that mobility offers more improvements in performance for individuals as returns (Holding et al., 2024; Tartari, Di Lorenzo, & Campbell, 2020). It is reported that individuals' social networks are supposed to be expanded since new connections will be set up as scientists move to another scientific affiliation while the previous connections will not disappear suddenly (Jiang, Pan, Wang, & Ma, 2024; Liu & Hu, 2022; Wang et al., 2019). Moreover, several studies have demonstrated scientific mobility could eventually improve individuals' performance in productivity and impact by comparing those moving scientists with their peers without moving experiences (Chen, Wu, Li, & Sun, 2023; Momeni et al., 2022; Uhlbach, Tartari, & Kongsted, 2022). The chances of collaboration, the probability of producing high-quality articles, and the internationalized impact are also discovered to be improved due to scientific mobility (Aykac, 2021; Gu et al., 2024).

Previous research highlighted the importance of scientific mobility. However, we supposed that the methods of mobility identification and performance evaluation are still at a coarse-grained level. As to mobility identifications, most studies depended on the changes in affiliated relationships to justify whether a focal scientist moved or not, and the information on affiliations is commonly extracted from published records (Aykac, 2021; Deville et al., 2014; Holding et al., 2024; Jiang et al., 2024; Liu & Hu, 2022; Momeni et al., 2022). Several studies also collected the mobility information by analyzing the author-level identifications, for instance, ORCID, Scopus ID, and Web of Science ID, or picking up affiliations information from individuals' curriculum vitae (CV) (Abramo et al., 2022; De Filippo et al., 2009; Tartari et al., 2020; Wang et al., 2019). Such methods might lack of strengths in interpreting how those short-term scientific activities, without changing affiliation information, could influence the scientists' performance in return. However, the gradually connective scientific communities and increasingly facilitating scholarly communications require demonstrations on whether short-term scientific activities, such as scientific visits, attending conferences, moving around for face-to-face collaboration, and conducting scientific experiments in another lab or facility abovementioned, will benefit or hurt scientists' performance. It is also a question attracting great attention from academia, policymakers, and the public.

Additionally, as to author-level performance evaluations, several studies took the mean value or positive probability of paper-level performance as a representation (Li, Tessone, & Zeng, 2024; Zeng, Fan, Di, Wang, & Havlin, 2022). However, we suppose that in the context of widespread collaborations, paper-level performance might need to be credited to co-authors respectively by measuring their contributions (Thelwall & Maflahi, 2022). Therefore, we introduced a cost-benefit perspective in this study and considered that all scientists' efforts during a specific period should be limited, dispersing to his or her scientific publications unevenly (Jones, 2021; Leyan Wu, Yi, Bu, Lu, & Huang, 2024). Therefore, the benefits of one scientist

attained from each publication depend on the costs he or she has invested (Zhang et al., 2024), and the volume of investment is measured by author sequence and based on the methods of proportional count (VanHooydonk, 1997).

Summary

Those short-term scientific activities without varying affiliations are named by us as Scientific Travels. They are increasingly common, but academia still knows little about scientific travels' impact on individuals' performance since, at the level of scientific big data, it is challenging to define and identify these activities with credit accuracy. However, the features of big science facility utilizations provide a valuable perspective and make such micro-identification possible. Based on the publications supported by worldwide big science facilities, the SLSs, it is easy to identify external scientists' global scientific activities and their flows during a specific period. Therefore, we are motivated to shrink this knowledge gap, provide important evidence on the impact of scientific travels, and support the decisions of science policy.

In the following sections, the analysis associates the travel experiences with scientists' performance, adjusted by individuals' contributions, and eventually offers a novel insight for related research in scientific mobility and enriches the framework of facilitymetrics.

Data & Method

Publication Library and Open Dataset

The scientific published data collection processes in the framework of Facilitymetrics are quite different since the special features of Big Science Facilities and should be noted. The traditional method, the retrieval query, was proved unsuitable due to lack of coverage and accuracy. If the published data were retrieved from Web of Science Core Collections (WoSCC) or Scopus, the fields of Affiliation Address and Funding Text should be applied. However, retrieving by Addresses might only lead to those publications at least authored by one staff who is affiliated with the focal facility while retrieving by Funding Text shall lead to those publications authored by external users, but the expressions of acknowledgments are not identical, and not all users acknowledged the focal facility in their publications (Silva et al., 2019; Söderström, 2023a, 2023b).

However, almost all Big Science Facilities around the globe have constructed their own bibliographic library to index their supporting scientific publications, and these libraries can be found and accessed on their official websites. Such libraries are considered one of the ways to make the scientific performance of big science facilities public and visible, responding to the concerns of policymakers, governments, and the public as taxpayers. Moreover, these libraries served an entrance for globally potential users to know the technological abilities and previous knowledge explored by the focal facility. Correspondingly, these libraries are considered self-constructed databases in the framework of Facilitymetrics, which highly facilitates the procedures of data collection.

We selected SLSs as our focal type of big science facilities in this work, a widely discussed type to be explored in Facilitymetrics, as abovementioned. SLSs are considered scientific platforms with advanced experimental technologies for almost all disciplines of science, especially material science, biology, physics, and chemistry. Concurrently, about 50 SLSs are operating or under construction around the world. Based on the expertise from China Big Science facilities and the guidance of the LightSources website[1], we constructed a publication dataset including about 240,000 scientific articles supported by 41 SLSs by exporting or crawling their self-constructed databases one by one. The remained 9 facilities have not constructed a mature database or have not been applied to support scientific research, and therefore, our dataset excluded them. For those collected facilities, not every SLS has operated for decades and possesses enough beamtime and experimental volume for global users. Therefore, in this study, we only considered the Top 20 SLSs (covered about 80% of publications) in productivity as analytical cases for better data quality. The selected big science facilities with their location, beginning year, and productivity (Final results after cleaning and matching with supplemental database by Python 3.11) are shown in Table 1.

Table 1. Selected Big Science Facilities (Top 20) and the Details of Publications.

No.	*Facility*	*Located Country/Region*	*Begin Year*	*Number of Publications*
1	ESRF	France	1986	26,544
2	APS	USA	1970	25,492
3	PETRA	Germany	1986	25,115
4	SPring-8	Japan	1999	12,922
5	ALS	USA	1991	12,733
6	PF	Japan	1972	11,091
7	Diamond	UK	2001	9,844
8	NSLS-II	USA	1984	9,005
9	SSRF	China	2000	8,207
10	MLS	Germany	1964	7,336
11	SSRL	USA	1983	5,731
12	AS	Australia	2006	5,659
13	NSRRC	Taiwan (China)	2003	5,629
14	BESSY	Germany	1992	5,621
15	PLS	Korea	2008	5,585
16	ELETTA	Italy	1994	5,182
17	NSRL	China	1984	4,821
18	LNLS	Brazil	1987	4,514
19	SOLEIL	France	2012	3,692
20	MAXIV	Sweden	1983	3,199
		Total Data		197,618

[1] https://lightsources.org/

It should be noted that every self-constructed database provides different structures of metadata, and the data framework is also differentiated, which highly challenges further data processing and limits our perspectives if we do not introduce bibliographic databases as supplemental data sources. Therefore, we used the OpenAlex database as a supplement to introduce more metadata by matching DOI and Title of published records collected from Top20 facilities' self-constructed databases. OpenAlex is a fully open database of the global research system with advantages in terms of inclusivity, affordability, and availability, and it is widely used in current research related to the science of science (Priem, Piwowar, & Orr, 2022).

Measures

We applied a 5-year Disruptive Index (DI_5) and Novelty Metrics, mainly Novelty Score (NS), as dependent variables to measure the scientific performance with the positive probabilities and Author Contribution (AC) adjusting mean value quantified. Moreover, we defined a new metric named the cutting-edge ability, which tells the boundaries-pushing by users' research to a focal facility by measuring the similarity with previous knowledge based on the Jaccard Similarity. Additionally, we set up a framework including several potential indicators to measure the correlations and regression relationships, for instance, the number of Traveled places, Traveled Times, resources of the network, and several involved knowledge topics. Details of our measurements are introduced as follows.

Scientific Performance

The Disruptive Index was proposed by Funk and Owen-Smith (2017) as CD-index and received a update by Lingfei Wu, Wang, and Evans (2019). It quantifies how one paper disrupts the current knowledge system according to the citation relationship. The illustration and formula are shown as follows:

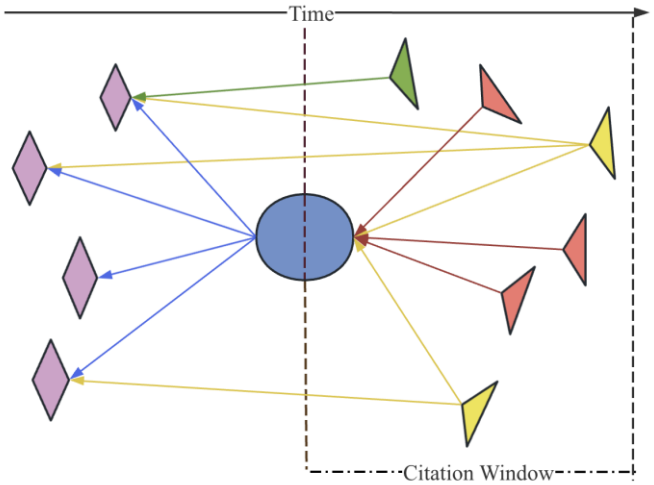


Figure 1. Illustration of Disruption Index.

$$(1)\text{Disruption Index} = \frac{N_r - N_y}{N_r + N_y + N_g}$$

For every focal paper (blue node in Figure 1), N_r represents the number of red triangles in Figure 1, measuring the citing publications that only cite the focal paper but do not cite its references, and the references of focal paper are displayed by purple rhombuses. N_y records the number of yellow triangles, telling those citing publications not only cite the focal paper but also cite its references while N_g means the number of citing publications that only cite the references of focal paper and colored in green in Figure 1. According to the formula, we could tell that the value should range from [-1, 1], and all red triangles lead to 1, indicating that the focal paper might create a new orientation in the current knowledge system, while all yellow triangles lead to -1, meaning that the focal paper might be a consolidative or developmental for its focal knowledge field. Therefore, if the value of DI was no less than zero, the focal paper was supposed to be disruptive. Otherwise, the focal paper was considered consolidating.

It is also obvious that DI might be influenced by the number of references, times cited, and the citation window. Therefore, we have set a 5-year citation window with at least five references and five citations as thresholds to ensure stability.

Novelty Metrics, consisting of Novelty Score and Conventionality Score, was proposed by Uzzi, Mukherjee, Stringer, and Jones (2013). It has introduced the concept of cited journal combinations to measure the focal paper's knowledge novel degree from the knowledge input perspective. The key step of the Novelty Score is the calculation of the Z-score, and the formula is shown as follows:

$$(2)\text{Z} = \frac{(\text{obs} - \text{exp})}{\sigma}$$

Every cited journal combination could be calculated a Z-score, and obs is the observed frequency of the focal cited journal pair while exp is the mean frequency of all cited journal pair and σ Represents the standard deviation of the number of journal pairs obtained from 10 randomized simulations of paper-to-paper citation network. Therefore, for one focal paper, its references and corresponding cited journal combinations could be found, and the Z-score of each combination could be sorted from the lowest to the highest, 10^{th} percentile Z-score is selected to represent the Novelty Score while the median Z-score is used to represent Conventionality Score.

Both Indicators, DI and NS, are widely explored and applied concurrently, and we applied them as two aspects of scientific performance to quantify the differences between scientific travelers and locals.

Author Contribution

We introduce a coefficient to adjust the evaluations of the author-level's scientific performance since this work mainly focuses on the scientific performance at the author-level (Zhang et al., 2024). We suppose that it is unsuitable to simply take the paper-level performance of an author in one specific year or during the total career

as his or her performance, especially concurrently, scientific collaborations are widespread, and scientists have a higher possibility to produce more than one papers in a year than before, leading to a situation that one scientist might distribute his or her efforts into several works simultaneously but unevenly. Therefore, we first filtered our data to retain those publications of teamwork and calculated the author contribution as an adjusting coefficient based on the method of proportional count and the hypothesis of cost-benefit perspective by measuring one author's rank in the team considered (VanHooydonk, 1997). The formula for Author Contribution is shown as follows:

$$(3)\text{Author Contribution} = \frac{(\mathrm{N}+1) - \mathrm{AS_a}}{\sum_1^{\mathrm{N}} \mathrm{AS}}$$

In formula (3), denoted N is the number of co-authors in one scientific team, while AS is the focal authors' sequence. If four authors collaboratively published one paper, the first author's credit should be 0.4, the last author's credit should be 0.1, and the two middle authors' credit should be 0.3 and 0.2, respectively. It is noted that this indicator is based on author sequence, which might overlook the contributions of corresponding authors of scientific teams. However, we suppose that the overlook might not cause heavy variations, and it is the most suitable choice. Firstly, the role of corresponding authors is difficult to identify in the level of publication data, and not all corresponding authors are always placed at the last. Moreover, corresponding authors usually have a higher tendency to publish more articles in one year or during the career than the first author and other authors, which well-matched our hypothesis that the efforts of the last author (if he or she is the corresponding author of his or her team) might be further distributed. If not, the last authors might be the lowest contribution author in the team.

We applied this coefficient to paper-level indicators of scientific performance and considered the mean values and positive probability of scientific performance adjusted by author contribution as the scientific performance of the focal author in one-year, total career, or for a specific time stage. The formulas of mean value and positive probability are as follows:

$$(4)\mathrm{Mean_j} = \frac{\sum_{\mathrm{i}}^{\mathrm{N}} (\mathrm{AC} \times \mathrm{P_i})}{\mathrm{N}}$$

$$(5)\mathrm{Prob_j} = \frac{\mathrm{N_{positive}}}{\mathrm{N}}$$

In formulas (4) and (5), denoted j is the period of scientific performance and AC is the focal author's credit in one paper and P_i is the scientific performance of corresponding paper. N should be the number of published articles of the focal author during the period j. $N_{positive}$ refers to the situation that $DI_5 \geq 0$ or $NS \leq 0$ and the probability does not need to be adjusted by author contribution since the sign will not change.

Traveled Places

The dataset of big science facilities' publications collected by us previously offers an even micro perspective to define the processes of scientific mobility since every facility requires users to conduct their experiments on-site. This context assists us in defining the role of scientific travelers and locals. We firstly confined that the focal authors should be external users of big science facilities, and if they have used more than one facility in a specific period, they should be scientific travelers. Otherwise, they are locals. The identification of the used facility is according to the relationships of focal author's publications with self-constructed databases. If one author's publication during a specific period is collected from more than one self-constructed database of facilities, we can tell that he or she should be a traveler since more than one facility is used. Therefore, the number of traveled places is considered as the number of used facilities in one year or during a specific period.

It should be noted that, according to our previous studies, the co-utilization between or among these big science facilities is uncommon but possible. Given that there is a co-utilized author who only published one publication but could be observed to use more than one facility. Such a situation is complicated and out of our research range, therefore, during the data cleaning, we have already dropped out those publications supported by more than one facility. It also means that Travelers should publish at least two articles in the focal period.

Other Important Indicators

We also define other indicators to finish further processes of visualization, correlations, and regression. Firstly, we proposed the volume of one author's network resources and involved knowledge topics from paper-level indicators by measuring the number of collaborative peers and published topics in a specific period. Secondly, we considered the productivity and the mean values of AC adjusted scientific impacts in one year and ten years to describe their impacts immediately and in the long term.

Furthermore, based on the Jaccard Similarity, we define the AC adjusted knowledge similarity by measuring the number of new topics in one publication compared with the using facilities' previously published topics numbers and considered the mean values to represent the performance of the focal author. The formula is shown as follows:

$$(6)\text{Knowledge Similarity} = \frac{\left|\overline{\mathrm{T}}_{\mathrm{i,j}} \cap \overline{\mathrm{T}}_{\mathrm{k,j-1}}\right|}{\left|\overline{\mathrm{T}}_{\mathrm{i,j}} \cup \overline{\mathrm{T}}_{\mathrm{k,j-1}}\right|}$$

Denoted paper i published in j year supported by facility k, and $\bar{T}_{i,j}$ refers to the research topics of focal paper while $\bar{T}_{k,j-1}$ refers to research topics the focal facility has researched. Both sets of topics are provided by OpenAlex. Then, the paper-level similarity with pervious knowledge could be calculated and after adjusting by AC, the mean values are used to indicate author-level performance during a specific period.

Additionally, we define the level of localization for travelers by measuring the ratio of local productivity and global productivity. Formulas are shown as follows:

$$(7-1)\text{Localization Ratio}_j = \frac{\text{Local Productivity}_j}{\text{Total Productivity}_j}$$

$$(7-2)\text{Divide Thresholds}_{n1} = \min_n \text{LR}_j + (\max_n \text{LR}_j - \min_n \text{LR}_j)/3$$

$$(7-3)\text{Divide Thresholds}_{n2} = \max_n \text{LR}_j - (\max_n \text{LR}_j - \min_n \text{LR}_j)/3$$

$$(7-4)\text{Localization Level} = \begin{cases} \text{Low}, & \min_n \text{LR}_j \le \text{LR}_j \le \text{DT}_{n1} \\ \text{Moderate}, & \text{DT}_{n1} < \text{LR}_j < \text{DT}_{n2} \\ \text{High}, & \text{DT}_{n2} \le \text{LR}_j \le \max_n \text{LR}_j \end{cases}$$

We first calculate the focal traveler's Localization Ratio in every used facility during the period j, and then find the lowest ratio and highest ratio of localization with the number of traveled facilities (denoted n in the formula 7-2 and 7-3) for all focal travelers during the period j considered. The divide thresholds could be found, and all focal travelers could be classified into different groups of Low, Moderate, and High according to the formula (7-4).

Results

We provided several perspectives related to the performance gap between scientific travelers and locals with multiple classifications applied to verify the robustness of our results. In the section of Results, we mainly classify external users into travelers or locals at the yearly level. The results by classifying at the level of total career or the level of past experiences are shown in the appendix, and all results are consistent, indicating the robustness of our discoveries and contributions. Moreover, the appendix also contains several figures for data distribution, which assisted us in setting thresholds for data filtering for better data quality.

Scientific Performance Gaps Between Travelers and Locals

According to Figure A1(A) and the definition of travelers abovementioned, the productivity of travelers and locals mainly distribute less than 15 articles, and therefore, we only considered those scientists' yearly productivity range from 2 to 15. From Figure A1(B), we can tell most scientists' career age is no more than 30 years, which leads to another threshold. Figure A1(C), displays the annual average credit differences between travelers and locals, and the value of author contribution is highly related to team size that we have confined that the number of co-authors in one article should be less than 45.

Figure A2 shows the tendency of modern science that the connections in global academia are increasingly close. As time goes on, more scientists tend to travel around, and concurrently the ratio of travelers in one year has reached 0.4. Figure A2(A) and A2(C) show similar results that middle-aged scientists have a higher possibility to travel to more than one facility, and junior scientists might lack travel chances.

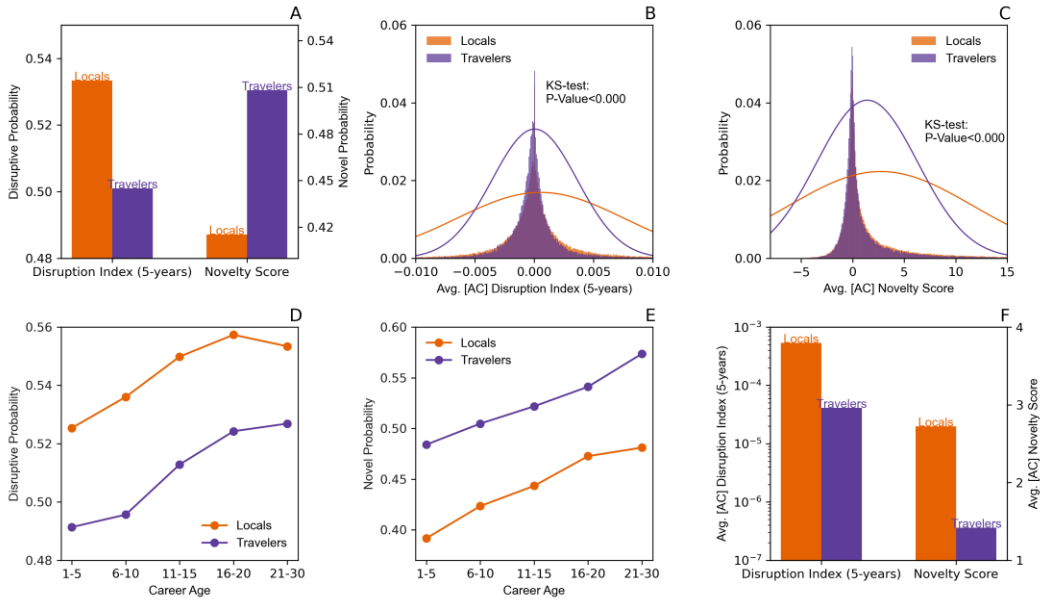


Figure 1. Travelers Associated with Better Novelty while Locals Produce More Disruption. Travelers and Locals are classified at the yearly level.

Under such context, Figure 1 mainly shows the basic results of this study that scientific travelers negatively related to disrupting the current knowledge systems while their works possess higher scientific novelty than locals. Figure 1(A) shows the gap of positive probability (K-S Test, $p<0.000$) between locals and travelers in scientific performance (103,359 Locals and 40,854 Travelers in the Sample of DI_5 while 142,420 Locals and 61,522 Travelers in the Sample of NS), and 1(B) and 1(C) display the mean value distribution of scientific performance indicators while 1(F) records the significant differences (K-S Test, $p<0.000$) of mean values between travelers and locals that locals still perform better at disruption but lack of knowledge novelty (Samples are consistent). 1(D) and 1(E) show the positive relationships between positive probability and career age.

Consistent results are also displayed in Figure A3 and Figure A4. Figure A3 classified all external users into "Never Traveled" and "Traveled" according to their travel experiences at career level, while Figure A4 identified "Un-Traveled" and "Over-Traveled" by yearly measuring whether the focal scientists have traveled or not in the past. For instance, given that there is one user (U) and he or she first traveled in 2000, leading to he or she is considered as an "Un-Traveled" before 2000, as an "Over-Traveled" current and after 2000. The results of the three classifications with their positive probabilities and mean values are consistent.

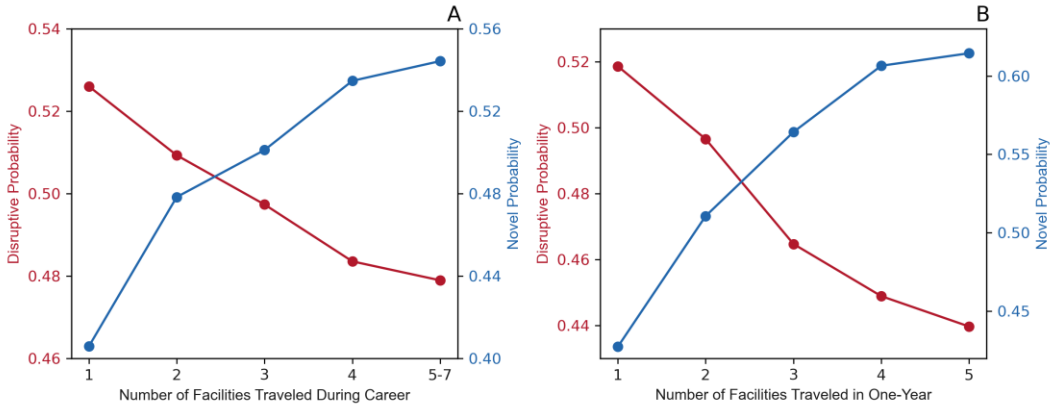


Figure 2. More Travels Lead to Negative Disruptive Ability but Positive Scientific Novelty.

Figure 2 displays relationships between the number of traveled facilities for scientists during their total career and in one year. The red color represents the variation of Disruptive Probability while the blue color shows the variation of Novelty. From the perspective of academic career, those locals might suffer from a low probability of novelty (about 0.4) but benefit from a high disruptive probability. The thresholds of traveled facilities numbers were selected by referring to Figure A5.

The Impacts of Localization

Denoted that the ratio of localization level describes the degree of concentration and dispersion of scientific travelers by measuring their local productivity and global productivity. If one traveler is observed with extremely skewed productivity in the minor facility, he or she might be a highly localized traveler. Here, we mainly classified scientists by their annual productivity, and the results of career-level productivity are shown in Figure A6.

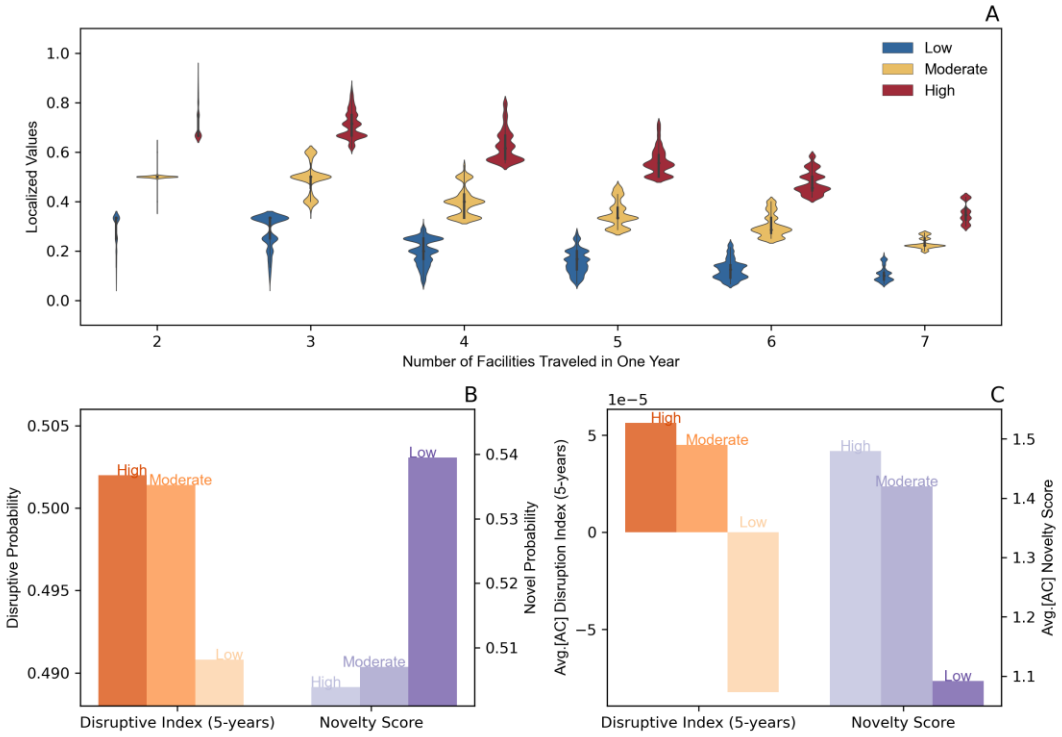


Figure 3. High Localization Travelers Produce More Disruption while Low Localization Travelers Associated with More Novelty.

Figure 3(A) displays the classifications of localization level with the number of traveled facilities considered. The threshold of traveled facilities numbers is also referred to in Figure A5. If one traveler's productivity ratio in any facility he or she used in the focal year drops in the red range, he or she is classified into high localization. Similarly, moderate and low levels of localization could be identified. In the sample of DI_5, 14,100 year-level Travelers are classified as High, 24,224 are classified as Moderate group, and 2,529 are classified as Low group. In the sample of NS, 20,418 year-level travelers are high localized, 37,023 are moderate, and 4,078 are low localized. In Figure 3(B) and Figure 3(C), results indicate that high localized travelers are associated with better disruptive performance than low localized counterparts while opposite results of novelty score. The performance gaps between different levels of localization are significant according to the K-S Test ($p<0.000$ for High-Moderate and Moderate-Low test when considering the positive probabilities of DI_5 and NS and the mean value of NS; $p<0.1$ for Moderate-Low test and $p<0.05$ for High-Moderate test when considering the mean values of DI_5). Figure A6 shows similar results that those scientists who traveled to several facilities during their career but have extremely skewed preferences might produce more disruptive knowledge while those who are not skewed in productivity might produce more novel knowledge. These two figures record the performance gap between travelers and if we take corresponding locals as controls to compare with, results support that

highly localized scientists produce less disruption knowledge but still better novelty than totally localized scientists.

The Impacts of Travel Experiences

In this subsection, we mainly focus on those scientists with travel experiences, and for the sake of improving the inclusive, we also included those travelers who have already localized and annually produced only one article in the local year.

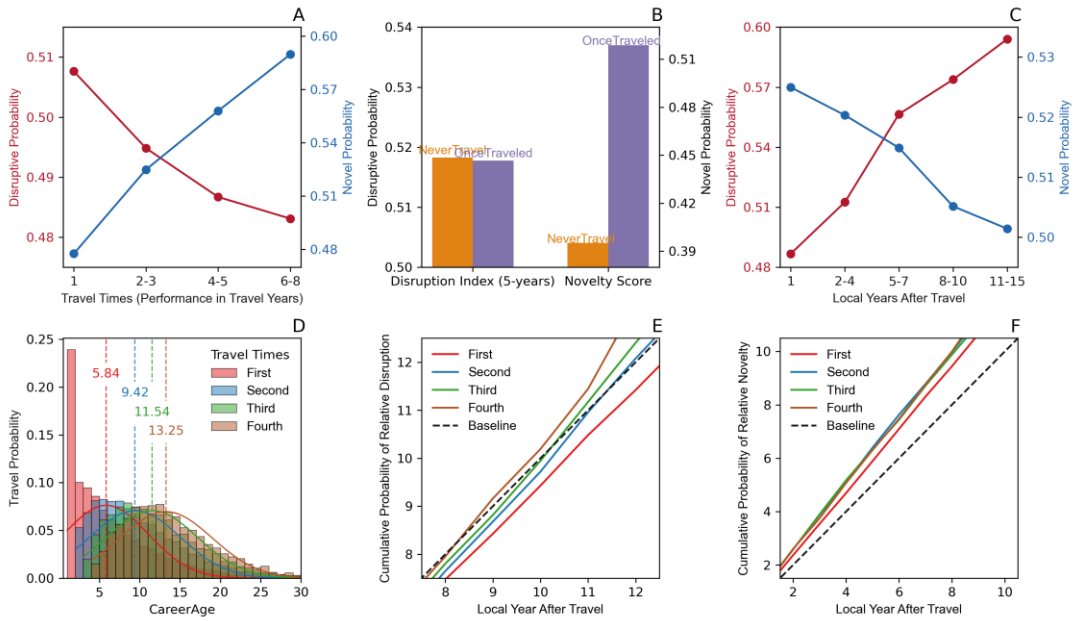


Figure 4. Travel Experience Damage Locals' Disruptive Ability but Increase Novelty.

Figure 4 displays evidence to understand how travel experiences will affect scientists' scientific performance. Firstly, Figure 4(A) shows the concurrent yearly performance variation of travelers in the travel year, indicating that more traveled experiences might decrease the probability of disruption but increase the novelty probability. Comparing the career-level performance of those scientific travelers when they are locals (Never Traveled to another facility, 62,480 year-level scientists for DI_5 and 65,997 for NS) and once traveled (at least traveled to two facilities previously, 42,114 year-level scientists for DI_5 and 75,717 for NS), and the results of comparisons are recorded in Figure 4(B). It is shown that the probability of disruption suffers from slight damage (KS-test: $p<0.000$, T-test: $p=0.854$) while novelty probability is observed a significant improvement (KS-test: $p<0.000$). The following figures could assist in understanding such a situation in Figure 4(C), we observed that for those travelers, once they have finished a one-year travel and are back to local scientists, their disruptive probabilities will increase as the local year goes on, but their probability of novelty might slowly decrease since total localization. However, the novel ability of these fully localized travelers is still much better than that of those locals without travel experiences. To better display the

variations, ensure data quality, and compare with those travelers' counterparts, we mainly focus on the impact of the first four times travel experience and display travelers' average travel career age as shown in Figure 4(D). Later, we take these mean values as the representative travel career ages by rounding down, considering the next year should be the first local year of those travelers who finished scientific travel, and select those locals with identical career ages as the control group to compare the subsequent years' performance whether a scientist chose to travel or not. Results are shown in Figure 4(E) and Figure 4(F), with the cumulative probability of relative disruption and novelty visualized. We consider those corresponding years' performance of locals as a baseline and compare it with the travelers' yearly scientific performance after their travels at different times. Then, the relative probability of positive scientific performance could be calculated, and eventually, the cumulative value could be found. From the abovementioned results, it is reported that scientific travel might decrease scientists' disruptive ability, and their disruption might increase gradually as they localized. However, Figure 4(E) argues that those scientists with travel experiences might slowly surpass their peers without travel experiences in disruptive ability as time goes on, especially those scientists with more than one-time travel experience, and the surpass year will become earlier if one traveler has traveled around for times. Figure 4(F) indicates that those scientists with travel experiences could significantly outperform their peers in producing novelty knowledge.

Alternative Indicators Differences between Travelers and Locals

Several factors might affect the performance gaps between travelers and locals with respect to previous knowledge. We aim to shrink such potential impacts and validate our results. Therefore, we visualized differences between travelers and locals in alternative indicators.

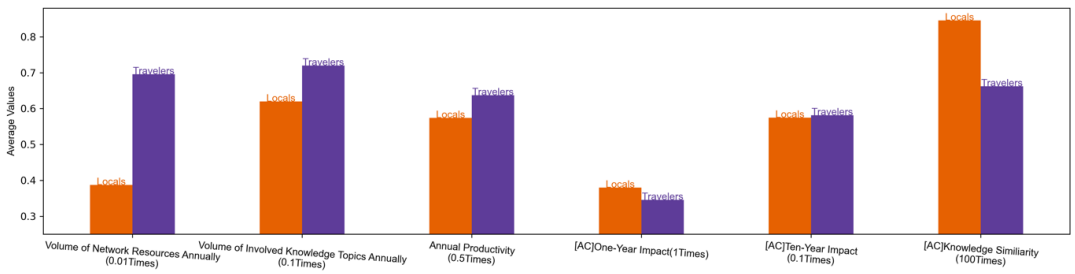


Figure 5. Alternative Indicators Differences Between Locals and Travelers.

Figure 5 tells the scientific input gaps between travelers and locals in annual network resources and involved knowledge topics and displays the output-level gaps in productivity, short-term and long-term impact, and the similarity with previous local knowledge. The values are normalized by us to reach a better visualization, and the times of normalizations are recorded following all indicators in the Figure. Locals might receive more short-term citations, while in the long term, travelers might have higher scientific impacts. Travelers might also perform better in expanding the

knowledge edge for the facility they are using since they have lower similarity with previous knowledge than locals.

Note that the annual volume of network resources represents the number of collaborators for a focal scientist in one year, and the annual volume of involved knowledge topics records the number of research topics the focal scientists have published. Both are reported to affect the scientific performance at the paper level and therefore, we put emphasis on them to avoid potential impacts on author-level performance and the results are shown in Figure 6.

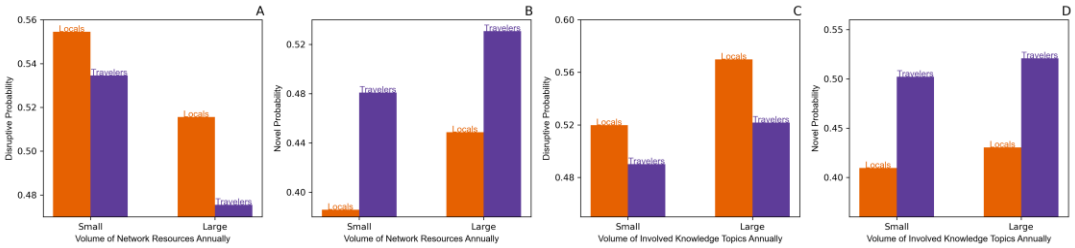


Figure 6. The Effects of Network and Involved Knowledge Topics on Scientific Performances.

Figure 6(A) and Figure 6(B) record the impact of small or large volumes of network resources on the scientific performance of travelers and locals, respectively. The classifications of small or large volumes refer to the distributions shown in Figure A7(A), and we take mean values (locals: 12 and travelers: 16) as boundaries. Even though a large volume of network resources might influence disruptive ability negatively and positively related to novel knowledge, the performance gaps between locals and travelers could still be observed that locals perform better in disruption while travelers could produce more novel knowledge. Similar tendencies could be discovered in Figure 6(C) and Figure 6(D) that if we control the impacts of involved knowledge topics (boundaries could be referred to in Figure A6(B)), locals still perform better in disruption, and travelers possess advantages in novelty.

Regression Analysis to validate

To validate the main results of this study, we conduct the Paper-level and author-level Ordinary Least Squares (OLS) regression to ensure the impacts of scientific travels on scientific performance. Table 2 displays the paper-level results with two corresponding indicators considered as independent variables respectively (the ratio of travelers and the total contribution of travelers in the focal academic team) and potentially influential variables controlled.

Specifically, we select Team Size (at least two co-authors), Number of References (at least five references), and Cited Topics as control variables for disruption index and novelty score according to our previous visualizations. Times Cited_5, a widely demonstrated impactful indicator on DI_5, is considered a unique control variable for disruptive index with a five-year citation window and at least five citations confined while the published year is customed for novelty score since the ability to advance knowledge might be affected by the level of scientific development. Moreover, we

consider the supporting facility of each publication as a dummy variable to avoid potential influence caused by different levels among technologies.

In the paper level, Table 2 demonstrates the negative impact of Travelers participating in the scientific team on disruptive ability as their ratio (β=-0.007, $p<0.001$) or contribution (β=-0.006, $p<0.001$) improving. The results in Table 2 also ensure the positive effects of Travelers on producing more novel knowledge, given that the lower value of Novelty Score represents better Novelty, and the increasing ratio and contribution of travelers could significantly improve research novelty. All regression models are significant according to F-scores and corresponding significances.

Table 2. Paper-level OLS regression with Indicators Related to Travelers in Teams Considered as Independent Variables.

Models	*(1)* DI_5	*(2)* DI_5	*(3)* *NS*	*(4)* *NS*
Travelers Ratio	-0.007***		-17.229***	
	(0.001)		(0.790)	
Travelers Contribution		-0.006***		-16.007***
		(0.001)		(0.815)
Team Size	-0.000***	-0.000***	-0.029	-0.032
	(0.000)	(0.000)	(0.037)	(0.037)
Number of References	-0.000***	-0.000***	-0.125***	-0.127***
	(0.000)	(0.000)	(0.009)	(0.009)
Cited Topics	0.000***	0.000***	-0.292***	-0.290***
	(0.000)	(0.000)	(0.012)	(0.012)
Times $Cited_5$	0.000***	0.000***		
	(0.000)	(0.000)		
Published Year			0.153***	0.139***
			(0.029)	(0.029)
Constant	0.008***	0.008***	-271.957***	-245.150***
	(0.000)	(0.000)	(57.477)	(57.435)
Dummy		Big science facility		
Adj. R^2	0.064	0.064	0.036	0.036
F-score	208.9***	207.3***	156.3***	152.5***
Obs.	72,896	72,896	99,425	99,425

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Author-level regressions could help to understand how scientific travels will influence scientific performance, as shown in Table A1 and Table A2. Both tables record the OLS regression results in the author-level performance evaluations but different from yearly and career perspectives, respectively, with a binary variable (Travelers: 1, Locals: 0) considered as the independent variable and career age, productivity, network resources, and involved topics controlled according to previous visualizations.

In Table A1, we conduct a yearly analysis which is consistent with our main figures, and results demonstrated that travelers negatively related to disruptive knowledge (β=-0.022, $p<0.001$ in $ProDI_5$ and β=-0.041, $p<0.001$ in $MeanDI_5$) but positively related to novel knowledge (β=0.037, $p<0.001$ in ProNS and β=-1.107, $p<0.001$ in MeanNS). Table A2 applied a career perspective that validate the robustness of our result that travelers still disadvantage in producing disruptive knowledge (β=-0.024, $p<0.001$ in $ProDI_5$ and β=-0.051, $p<0.001$ in $MeanDI_5$) but associated with more novelty (β=0.044, $p<0.001$ in ProNS and β=-1.544, $p<0.001$ in MeanNS).

Discussion and Conclusion

This study provides a more micro and, therefore, more novel perspective to identify the impacts of short-term scientific travels on individuals' scientific performance, quantified by disruptive index and novelty score, discovered that travelers might disturb scientists' ability to produce disruptive knowledge but enhance their novelty ability in return. The micro identification is beneficial from the features of utilizing big science facilities, mainly the characters of external users and on-site experiments. Results classified two types of external users (travelers and locals) by multi-approaches from yearly, previous, and career perspectives, and all results are consistent to show locals associated with higher disruption while travelers perform better in novel knowledge production. Further results indicate that the performance loss of travelers in disruption is mainly short-term, and the last of the period averagely depends on their travel times. We observed that their disruptive ability might increase and even surpass those peers without travel experiences since they have finished their scientific travels and become local users as time goes by. The novel abilities of Travelers are observed to be significantly higher than those of locals in different classifications. Additionally, we conduct OLS regressions at the paper level and author level, respectively, to validate the robustness and consistency of our results. The results of causal inference provided further evidence to support our main conclusions.

The micro-level identification has enriched the extant research in scientific mobility associated with scientific performance since our methods make use of the features in big science facilities context, and the results could be extrapolated to similar situations such as visiting scholars, attending conferences, and any other activities for scientific communication and collaboration without affiliated information to be identified. After all, we propose positive evidence to those policies encouraging scientific mobility and scientific communication, and we demonstrate that in long-term scientific' careers, those travelers could produce novel knowledge easier than those scientists without travel experiences but insignificantly suffer from the loss of disruptive ability.

This study also has several limitations. Firstly, the loss of data should be noted, and the volume of published records is limited by the operating years and experimental volumes of big science facilities for external users. The process of data collection also receives lots of challenges due to one facility having one customed database, and some of them provide low-quality publication data. Therefore, we only take about 210,000 articles as the sample, which might shrink the applied scope of results.

Secondly, concurrently, most advanced facilities located in developed countries or regions and open to their citizens might be the priorities, leading to the scientific contributions from global south might be overlooked potentially. We highly recommend future research focusing on related issues and providing more solutions.

Acknowledgments

This work was supported by the National Social Science Fund Major Projects of China (Project No. 22&ZD127). We would like to thank Xiaowei ZHANG, Yuhui DONG, and Honghong LI for their expertise in big science facilities. We also appreciate the constructive comments from reviewers.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2022). The effect of academic mobility on research performance: The case of Italy. *Quantitative Science Studies, 3*(2), 345-362. doi:10.1162/qss_a_00192

Aykac, G. (2021). The value of an overseas research trip. *Scientometrics, 126*(8), 7097-7122. doi:10.1007/s11192-021-04052-4

Chen, Y. T., Wu, K. Y., Li, Y., & Sun, J. J. (2023). Impacts of inter-institutional mobility on scientific performance from research capital and social capital perspectives. *Scientometrics, 128*(6), 3473-3506. doi:10.1007/s11192-023-04690-w

Conroy, G. (2024). World's brightest X-rays: China first in Asia to build next-generation synchrotron. *Nature, 629*(8013), 740. doi:10.1038/d41586-024-01346-4

D'Ippolito, B., & Rüling, C. C. (2019). Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications. *Research Policy, 48*(5), 1282-1296. doi:10.1016/j.respol.2019.01.011

De Filippo, D., Casado, E. S., & Gómez, I. (2009). Quantitative and qualitative approaches, to the study of mobility and scientific performance: a case study of a Spanish university. *Research Evaluation, 18*(3), 191-200. doi:10.3152/095820209x451032

Deville, P., Wang, D. S., Sinatra, R., Song, C. M., Blondel, V. D., & Barabási, A. L. (2014). Career on the Move: Geography, Stratification, and Scientific Impact. *Scientific Reports, 4*, 7. doi:10.1038/srep04770

Franzoni, C., Scellato, G., & Stephan, P. (2012). Foreign-born scientists: mobility patterns for 16 countries. *Nature Biotechnology, 30*(12), 1250-1253. doi:10.1038/nbt.2449

Funk, R. J., & Owen-Smith, J. (2017). A Dynamic Network Measure of Technological Change. *Management Science, 63*(3), 791-817. doi:10.1287/mnsc.2015.2366

Gu, J. W., Pan, X. L., Zhang, S. X., & Chen, J. Y. (2024). International mobility matters: Research collaboration and scientific productivity. *Journal of Informetrics, 18*(2), 15. doi:10.1016/j.joi.2024.101522

Hallonsten, O. (2013). Introducing 'facilitymetrics': a first review and analysis of commonly used measures of scientific leadership among synchrotron radiation facilities worldwide. *Scientometrics, 96*(2), 497-513. doi:10.1007/s11192-012-0945-9

Hallonsten, O. (2014). How expensive is Big Science? Consequences of using simple publication counts in performance assessment of large scientific facilities. *Scientometrics, 100*(2), 483-496. doi:10.1007/s11192-014-1249-z

Hallonsten, O. (2016). Use and productivity of contemporary, multidisciplinary Big Science. *Research Evaluation, 25*(4), 486-495. doi:10.1093/reseval/rvw019

Hallonsten, O., & Christensson, O. (2017). Collaborative technological innovation in an academic, user-oriented Big Science facility. *Industry and Higher Education, 31*(6), 399-408. doi:10.1177/0950422217729284

Hand, E. (2010). 'Big science' spurs collaborative trend. *Nature, - 463*(- 7279), - 282. Retrieved from - https://doi.org/10.1038/463282a

Heidler, R., & Hallonsten, O. (2015). Qualifying the performance evaluation of Big Science beyond productivity, impact and costs. *Scientometrics, 104*(1), 295-312. doi:10.1007/s11192-015-1577-7

Heinze, T., & Hallonsten, O. (2017). The reinvention of the SLAC National Accelerator Laboratory, 1992-2012. *History and Technology, 33*(3), 300-332. doi:10.1080/07341512.2018.1449711

Holding, B. C., Acciai, C., Schneider, J. W., & Nielsen, M. W. (2024). Quantifying the mover's advantage: transatlantic migration, employment prestige, and scientific performance. *Higher Education, 87*(6), 1749-1767. doi:10.1007/s10734-023-01089-7

Jiang, F., Pan, T. X., Wang, J., & Ma, Y. F. (2024). To academia or industry: Mobility and impact on ACM fellows' scientific careers. *Information Processing & Management, 61*(4), 15. doi:10.1016/j.ipm.2024.103736

Jiménez, C. (2010). Synching Europe's big science facilities. *Nature, - 464*(- 7289), - 659. Retrieved from - https://doi.org/10.1038/464659a

Jones, B. F. (2021). The Rise of Research Teams: Benefits and Costs in Economics. *Journal of Economic Perspectives, 35*(2), 191-216. doi:10.1257/jep.35.2.191

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1-18. doi:10.1016/s0048-7333(96)00917-1

Lauto, G., & Valentin, F. (2013). How Large-Scale Research Facilities Connect to Global Research. *Review of Policy Research, 30*(4), 381-408. doi:10.1111/ropr.12027

Li, H. Y., Tessone, C. J., & Zeng, A. (2024). Productive scientists are associated with lower disruption in scientific publishing. *Proceedings of the National Academy of Sciences of the United States of America, 121*(21), 9. doi:10.1073/pnas.2322462121

Lin, Y., Frey, C. B., & Wu, L. (2023). Remote collaboration fuses fewer breakthrough ideas. *Nature, 623*(7989), 987-991. doi:10.1038/s41586-023-06767-1

Liu, M. J., & Hu, X. (2022). Movers? advantages: The effect of mobility on scientists? productivity and collaboration. *Journal of Informetrics, 16*(3), 17. doi:10.1016/j.joi.2022.101311

Momeni, F., Karimi, F., Mayr, P., Peters, I., & Dietze, S. (2022). The many facets of academic mobility and its impact on scholars'. *Journal of Informetrics, 16*(2), 19. doi:10.1016/j.joi.2022.101280

Priem, J., Piwowar, H. A., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv, abs/2205.01833*.

Silva, F. S. V., Schulz, P. A., & Noyons, E. C. M. (2019). Co-authorship networks and research impact in large research facilities: benchmarking internal reports and bibliometric databases. *Scientometrics, 118*(1), 93-108. doi:10.1007/s11192-018-2967-4

Söderström, K. R. (2023a). Global reach, regional strength: Spatial patterns of a big science facility. *Journal of the Association for Information Science and Technology, 74*(9), 1140-1156. doi:10.1002/asi.24811

Söderström, K. R. (2023b). The structure and dynamics of instrument collaboration networks. *Scientometrics, 128*(6), 3581-3600. doi:10.1007/s11192-023-04658-w

Tartari, V., Di Lorenzo, F., & Campbell, B. A. (2020). "Another roof, another proof": the impact of mobility on individual productivity in science. *Journal of Technology Transfer, 45*(1), 276-303. doi:10.1007/s10961-018-9681-5

Thelwall, M., & Maflahi, N. (2022). Research coauthorship 1900-2020: Continuous, universal, and ongoing expansion. *Quantitative Science Studies, 3*(2), 331-344. doi:10.1162/qss_a_00188

Uhlbach, W. H., Tartari, V., & Kongsted, H. C. (2022). Beyond scientific excellence: International mobility and the entrepreneurial activities of academic scientists. *Research Policy, 51*(1), 16. doi:10.1016/j.respol.2021.104401

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science, 342*(6157), 468-472. doi:10.1126/science.1240474

van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology, 58*, 515-541. doi:10.1146/annurev.psych.58.110405.085546

Van Noorden, R. (2012). SCIENCE ON THE MOVE. *Nature, 490*(7420), 326-329. doi:10.1038/490326a

VanHooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science, 48*(10), 944-945. doi:10.1002/(sici)1097-4571(199710)48:10<944::Aid-asi8>3.0.Co;2-1

Verginer, L., & Riccaboni, M. (2021). Talent goes to global cities: The world network of scientists' mobility. *Research Policy, 50*(1), 17. doi:10.1016/j.respol.2020.104127

Wang, J., Hooi, R., Li, A. X., & Chou, M. H. (2019). Collaboration patterns of mobile academics: The impact of international mobility. *Science and Public Policy, 46*(3), 450-462. doi:10.1093/scipol/scy073

Wild, S. (2021). Plan for Africa's first synchrotron light source starts to crystallize. *Nature*. doi:10.1038/d41586-021-02938-0

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378-382. doi:10.1038/s41586-019-0941-9

Wu, L., Yi, F., Bu, Y., Lu, W., & Huang, Y. (2024). Toward scientific collaboration: A cost-benefit perspective. *Research Policy, 53*(2), 104943. doi:https://doi.org/10.1016/j.respol.2023.104943

Xu, H. M., Liu, M. J., Bu, Y., Sun, S. J., Zhang, Y., Zhang, C. W., . . . Ding, Y. (2024). The impact of heterogeneous shared leadership in scientific teams. *Information Processing & Management, 61*(1), 13. doi:10.1016/j.ipm.2023.103542

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences of the United States of America, 119*(36), 8. doi:10.1073/pnas.2200841119

Zeng, A., Fan, Y., Di, Z. G., Wang, Y. G., & Havlin, S. (2022). Impactful scientists have higher tendency to involve collaborators in new topics. *Proceedings of the National Academy of Sciences of the United States of America, 119*(33), 9. doi:10.1073/pnas.2207436119

Zhang, M.-Z., Wang, T.-R., Lyu, P.-H., Chen, Q.-M., Li, Z.-X., & Ngai, E. W. T. (2024). Impact of gender composition of academic teams on disruptive output. *Journal of Informetrics, 18*(2), 101520. doi:https://doi.org/10.1016/j.joi.2024.101520

Appendix

Table A1. OLS Regression of Scientific Performance at the Level of Publish Year.

Models	*(1)* *$ProDI_5$*	*(2)* *$MeanDI_5$*	*(3)* *ProNS*	*(4)* *MeanNS*
T1L0	-0.022***	-0.041***	0.037***	-1.107***
	(0.001)	(0.003)	(0.001)	(0.039)
Career Age of the year	0.003**	-0.020***	0.029***	-0.068***
	(0.001)	(0.003)	(0.001)	(0.003)
Annual Productivity	-0.066***	-0.072***	-0.031***	0.276***
	(0.002)	(0.005)	(0.002)	(0.017)
Annual Network Resources	0.009***	0.019***	-0.008***	0.000***
	(0.001)	(0.003)	(0.001)	(5.83e-05)
Annual Involved Topics	0.084***	0.095***	0.032***	-0.177***
	(0.002)	(0.005)	(0.001)	(0.010)
Constant	0.524***	0.000	0.443***	3.498***
	(0.001)	(0.003)	(0.001)	(0.042)
Adj. R^2	0.017	0.004	0.020	0.011
F-Score	486.6***	130.6***	854***	460.1***
Obs.	144,213	144,213	203,942	203,890

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All Independent variables are standardized to mean zero and S.E. 1, and for better discoveries in data, we also standardized the dependent variables of $MeanDI_5$. T1L0 is a binary variable that denoted Travelers as 1 while Locals as 0

Table A2. Robustness Check of Scientific Performance at the Level of Author Career Age.

Models	*(1)* *$ProDI_5$*	*(2)* *$MeanDI_5$*	*(3)* *ProNS*	*(4)* *MeanNS*
T1L0	-0.024***	-0.051***	0.044***	-1.544***
	(0.002)	(0.004)	(0.001)	(0.060)
Career Age	0.005**	0.005	-0.007***	-0.013*
	(0.002)	(0.004)	(0.001)	(0.007)
Total Productivity	-0.028***	-0.029***	-0.015***	0.018***
	(0.002)	(0.006)	(0.002)	(0.004)
Total Network Resources	-0.004**	0.012**	-0.015***	0.000***
	(0.001)	(0.004)	(0.001)	(2.15e-05)
Total Involved Topics	0.041***	0.048***	0.020***	-0.023***
	(0.002)	(0.006)	(0.002)	(0.003)
Constant	0.525***	0.000	0.430***	3.521***
	(0.001)	(0.004)	(0.001)	(0.044)
Adj. R^2	0.008	0.003	0.018	0.011
F-Score	110.1***	43.89***	327.8***	201.9***
Obs.	67,441	67,441	89,963	89,911

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All Independent variables are standardized to mean zero and S.E. 1, and for better discoveries in data,

we also standardized the dependent variables of $MeanDI_5$. T1L0 is a binary variable that denoted Travelers as 1 while Locals as 0

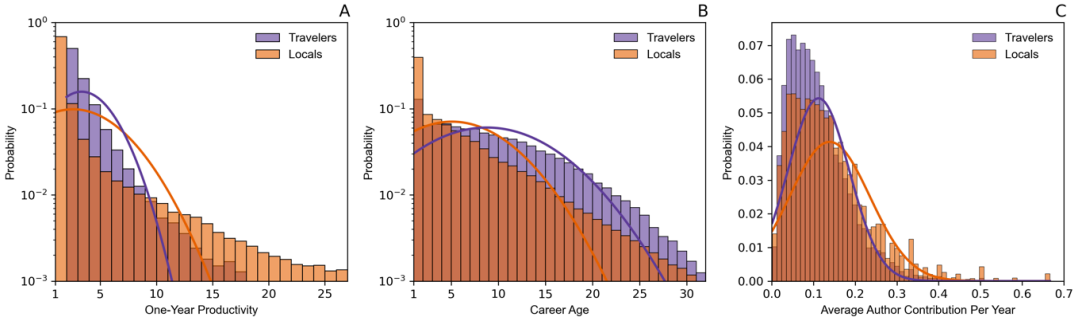


Figure A1. Probability of Travelers and Locals Yearly Productivity, Career Age, and Averagely Collaborative Contribution. Therefore, we selected those authors whose one-year productivity no more than 15, career age no more than 30 and limited the team size of published records less than 45 due to credits of Author Contribution.

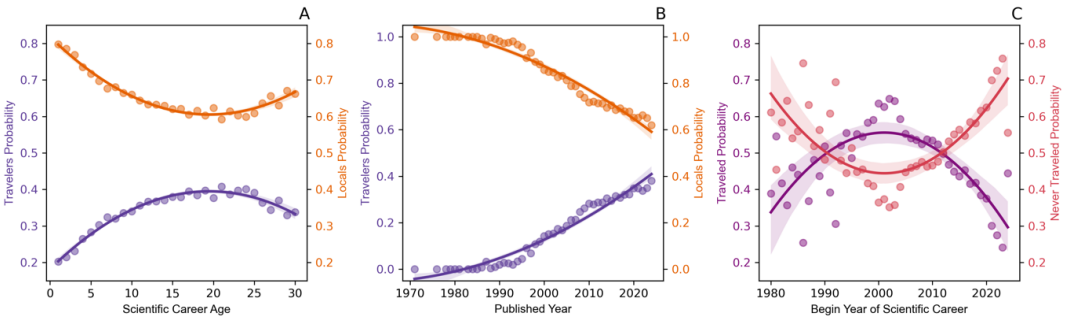


Figure A2. Probability of Traveler and Locals/Non-Travelers.

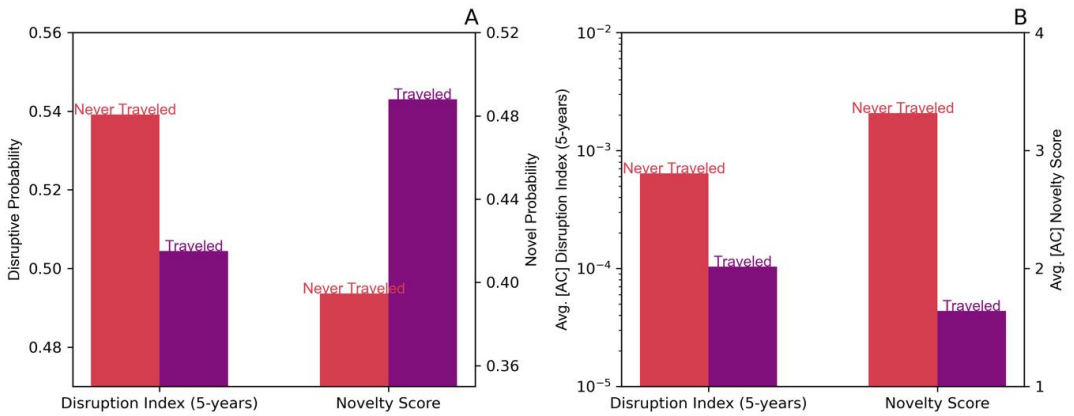


Figure A3. Identical Differences in Performance Between Locals and Travelers in Career Scale. Denoted "Never Traveled" represents those scientists who used only one facility during the career while "Traveled" means those scientists who used at least two facilities during the career.

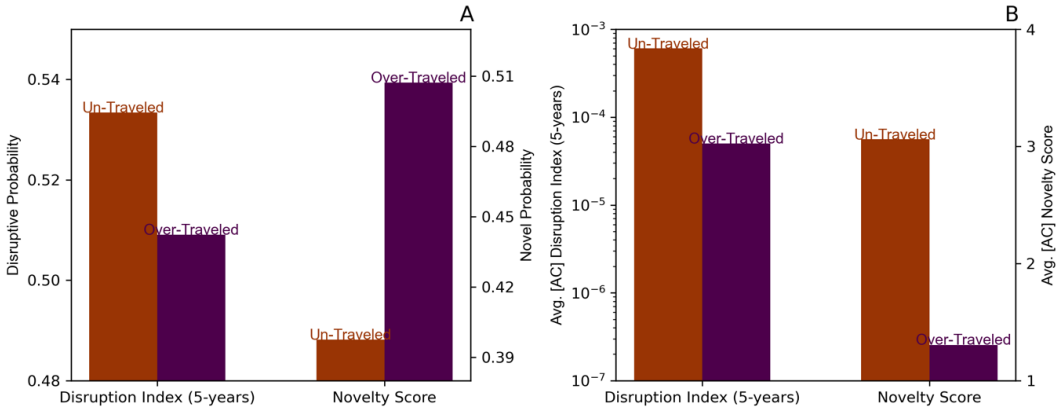


Figure A4. Identical Differences in Performance Once Scientific Travel Appeared. Denoted "Un-Traveled" represents those scientists have not used more facilities and "Over-Traveled" represents those scientists have used more facilities. Once the author used more than one facility, the author would be considered from "Un-Traveled" to "Over-Traveled".

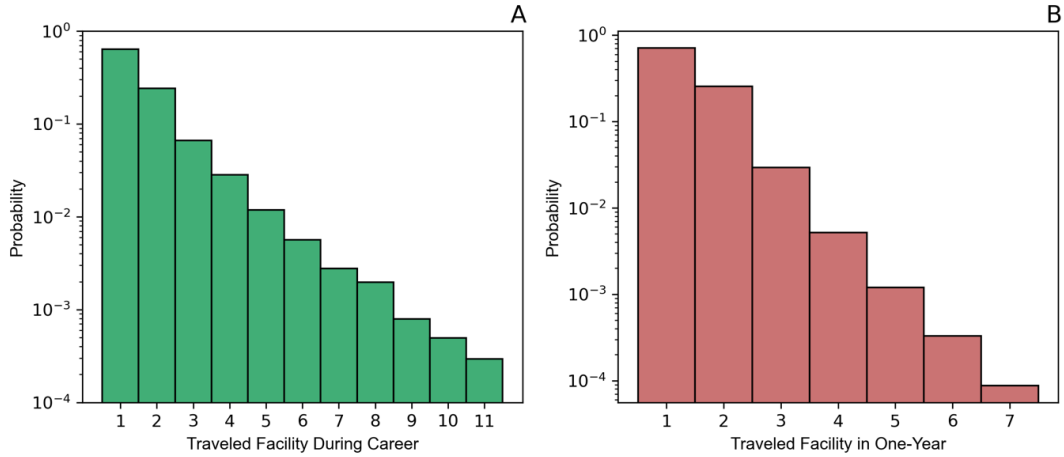


Figure A5. Probability of Traveled Facility Numbers for Scientists during Career and Yearly. We selected the thresholds as no more than seven facilities during career and no more than five facilities in One-Year.

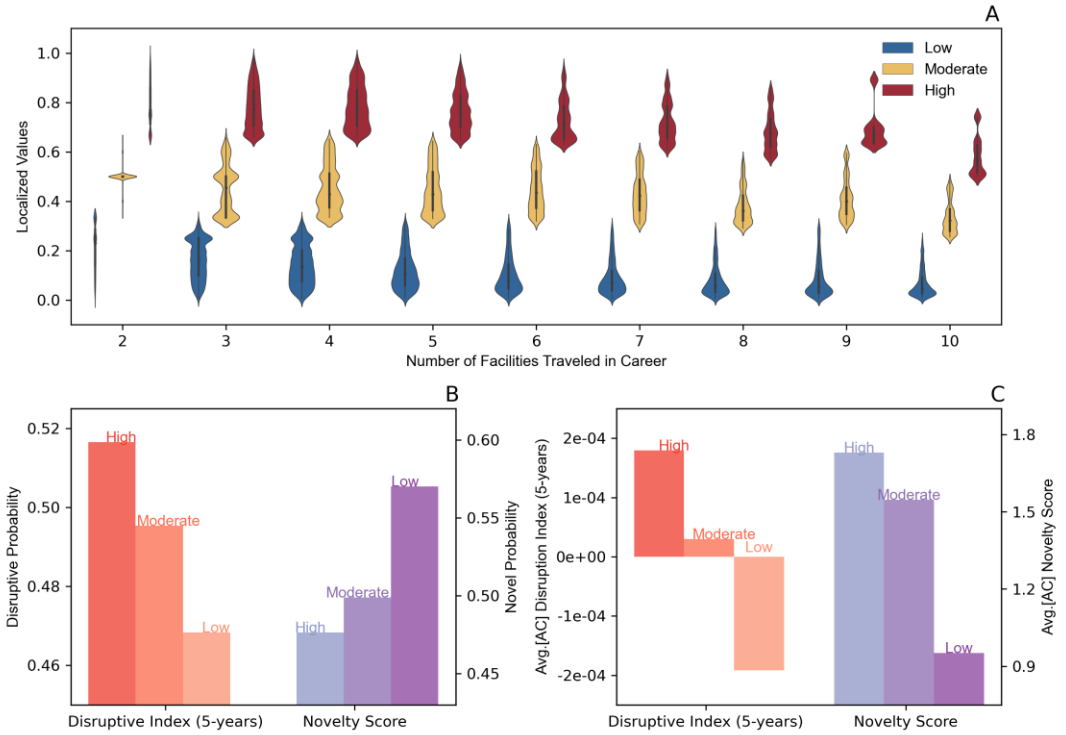


Figure A6. Performance of Travelers with Different Localization Levels in Career Scale. Performance gaps are consistent and more obvious in the career scale and high localized travelers associated with better disruption but lower novelty.

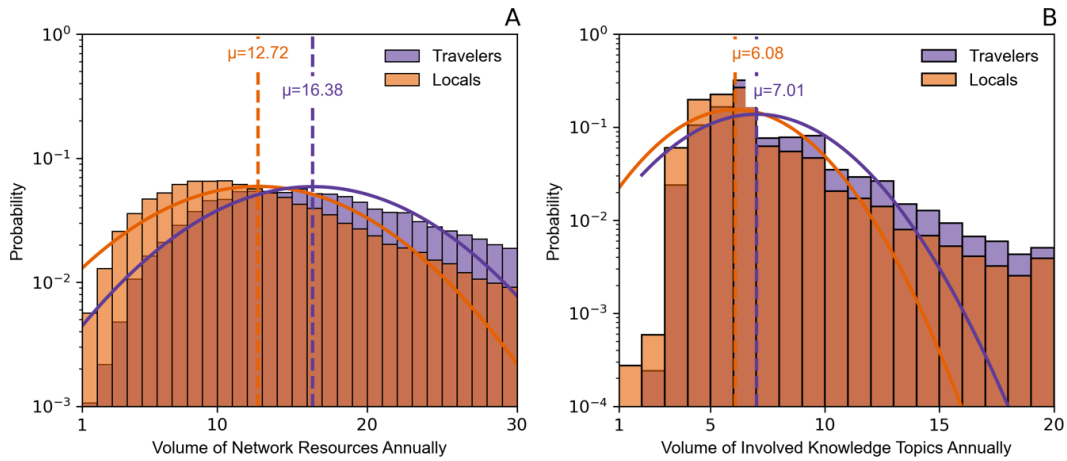


Figure A7. Performance of Travelers with Different Localization Levels in Career Scale. We considered 12 and 16 respectively for locals and travelers as thresholds to divide their volume of network resources. Six and Seven are respectively take as thresholds to divide the annually volume of involved knowledge topics for locals and travelers.

https://doi.org/10.51408/issi2025_033

Self Citations in Academic Excellence: Analysis of the Top 1% Highly Cited India-Affiliated Research Papers

Kiran Sharma[1], Parul Khurana[2]

[1]*kiran.sharma@bmu.edu.in*
School of Engineering & Technology, BML Munjal University, Gurugram, Haryana-122413 (India)
Center for Advanced Data and Computational Science, BML Munjal University, Gurugram, Haryana-122413 (India)

[2]*parul.khurana@lpu.co.in*
School of Computer Applications, Lovely Professional University, Jalandhar - Delhi G.T. Road, Phagwara, Punjab – 144411 (India)

Abstract

Citations demonstrate the credibility, impact, and connection of a paper with the academic community. Self citations support research continuity, but, if excessive, may inflate metrics and raise bias concerns. The aim of the study is to examine the role of self citations towards the research impact of India. To study this, 3.58 million papers affiliated with India from 1947 to 2024 in the Scopus database were downloaded, and 2.96 million were filtered according to document type and publication year up to 2023. Further filtering based on high citation counts identified the top 1% of highly cited papers, totaling 29,556. The results indicate that the impact of Indian research, measured by highly cited papers, has grown exponentially since 2000, reaching a peak during the 2011–2020 decade. Among the citations received by these 29,556 papers, 6% are self citations. Papers with a high proportion of self citations (>90%) are predominantly from recent decades and are associated with smaller team sizes. The findings also reveal that smaller teams are primarily domestic, whereas larger teams are more likely to involve international collaborations. Domestic collaborations dominate smaller team sizes in terms of both self citations and publications, whereas international collaborations gain prominence as team sizes increase. The results indicate that while domestic collaborations produce a higher number of highly cited papers, international collaborations are more likely to generate self citations. The top international collaborators in highly cited papers are the USA, followed by UK, and Germany.

Introduction

Citations are essential in academic research, acknowledging previous work, demonstrating integrity, and situating new studies within a broader scientific context. They serve as a key metric for assessing the impact of research, with high citation counts reflecting significant contributions to the field. Furthermore, citations facilitate knowledge dissemination, foster collaboration, and link studies between disciplines (Bornmann and Daniel, 2008). However, self citations, where authors cite their own work, provide continuity by linking new findings to prior contributions, especially in cumulative research. They also increase the visibility of newly published papers, which can attract external citations by highlighting related work (Hyland, 2003). However, excessive self citations can artificially inflate the citation metrics, misrepresenting the true influence of the paper, and raising concerns about bias (Fowler and Aksnes, 2007). Thus, while citations and self citations are vital

tools for measuring academic impact, their appropriate use is essential to maintain credibility and transparency in research.

Moreover, when self citations are used in an excessive or strategic manner to inflate citation metrics, it distorts the author's as well as the organization's academic influence (Moed, 2006). Today, citations in the form of scientific influence are used by various academic and government organizations for hiring, promotions, institutional prestige, bridging knowledge across different fields, fostering interdisciplinary research and funding decisions (Van Leeuwen, 2013). In such cases, self citations may create citation loops to potentially skew critical bibliometric indicators (Taham- tan and Bornmann, 2019). The existence of groups in the form of "citation cartels" also engage in reciprocal citation practices, which further compounds the issue (Hillman and Baydoun, 2019). This trend underscores the urgent need for a nuanced understanding of self citations across different academic backgrounds as high citations indicate that the particular study has substantial contribution in the field of research (Hirsch, 2005).

As the scientific community understood the elevation, narrative, and opportunistic power of self citations, concerns arose about the ethical implications of artificial citations (Van Noorden and Chawla, 2019). Some argued in favor of self citations, stating it as a reflection of specialization, while others presented them as manipulations (Costas et al., 2010). The tipping point came when researchers unravel the self citation patterns at the level of authors, country exhibitions, and academic organizations (Hellsten et al., 2007).

Citations in the academic world work as the thread that weaves the vast fabric of human knowledge. If utilized properly and for the advancement of the community, they are more than just numbers (Hodge, 2025; Szomszor et al., 2020). They enhance human knowledge by guaranteeing coherence and continuity. Institutions and financial agencies need to take into account the caliber of contributions rather than just the quantity of citations (Hussein et al., 2024). Most significantly, the scientific community needs to keep improving its evaluation methods so that a researcher's effect is determined by academic merit rather than metric manipulation (Martin, 2013).

Research objectives

The study examines the top 1% of highly cited papers and aims to achieve the following objectives:

1. Evaluate the concentration of self citations in the highly cited papers over year and decades.
2. Investigate the influence of team size on self citation patterns.
3. Explore the impact of domestic and international collaborations on highly cited papers and the associated concentration of self citations.

Methodology

In figure 1, the flow chart outlines the process of selecting the top 1% highly cited research papers affiliated with India, based on data retrieved from Scopus. A total of 3.58 million papers from 1947 to 2024 were downloaded from Scopus, searching for

the affiliation country as "India". The dataset was then refined to include only articles, conference proceedings, and reviews, focusing on publications up to 2023, resulting in 2.96 million papers from 1947 to 2023. Of these, articles accounted for 2.28 million (76.29%), conference proceedings for 0.56 million (18.94%) and reviews for 0.14 million (4.78%). The papers were organized in descending order based on the number of citations received. Further filtering identified the top 1% of highly cited papers, totalling 29,556. Within this group, articles comprised 21,645 papers (73.23%), reviews 7,100 (24%) and conference papers 811 (2.74%). Finally, the filtering process systematically narrowed down a massive dataset of more than 3.5 million papers to a smaller set of highly impactful publications. The top 1% highly cited papers (29,556) represent the most influential research outputs affiliated with India, highlighting the country's global academic and scientific contributions.

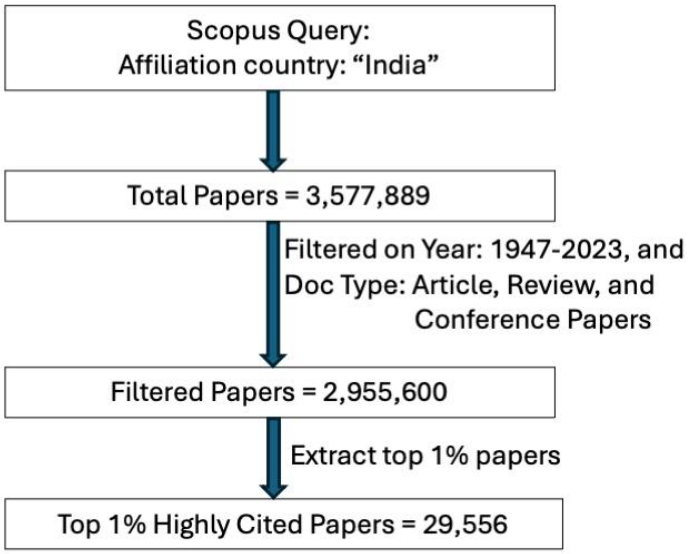


Figure 1. Search strategies undertaken to identify top 1% highly cited papers affiliated to India.

Results and Discussion

Citations vs. self citations

Citations and self citations are indispensable tools for academic research, helping to recognize prior work, measure impact, and foster scholarly communication. Striking the right balance between self-referencing and engaging with the broader academic community is essential to maintain the integrity and quality of research. Figure 2 represents the trends in total citations and number of papers (logarithmic scale) affiliated with India over time (from 1947 to 2023). Very few highly cited papers were published during 1947–1980 (early stage), as reflected by the flat portion of the blue star line. Citations are also minimal, but the red dotted line shows occasional spikes (possibly due to a few influential papers published during this period). The

graph demonstrates India's remarkable progress in producing highly influential research papers, particularly post-2000.

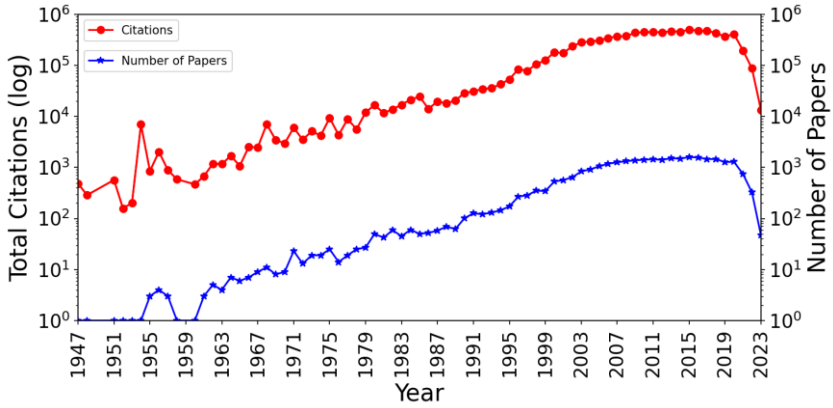


Figure 2. Year-wise trend of number of publications and total citations received.

A total of 29,556 highly cited papers received 97,53,620 citations in total, as shown in Table 1. Of these 9.75 million citations, 0.59 million (593,321) are self-cited, that is, 6% citations are self citations. Figure 3 represents the trend line of the average self citations over years received by the highly cited papers. The data exhibit fluctuations, with periods of increase and occasional declines, but the overall trend reveals a steady growth in self citations over time. The positive slope of the red trend line confirms this upward trajectory, suggesting that self citations have generally become more frequent in recent years.

Table 1. Search strategies undertaken to identify top 1% highly cited papers affiliated to India.

Decades	TP		Total Citations	Self Citations	Open Access	
	Count	%Count			Yes	No
1947-1950	2	0.01	801	0	0	2
1951-1960	15	0.05	13,083	26	1	14
1961-1970	61	0.21	24,585	364	3	58
1971-1980	193	0.65	76,997	2,325	8	185
1981-1990	547	1.85	1,92,674	8,658	34	513
1991-2000	2,042	6.91	7,88,617	45,757	175	1,867
2001-2010	9,680	32.75	33,79,239	1,85,869	1,204	8,476
2011-2020	14,614	49.45	48,98,374	3,25,447	3,929	10,685
2021-2023	2,402	8.13	3,79,250	24,875	973	1,429
Total	**29,556**	**100**	**97,53,620**	**5,93,321**	**6,327**	**23,229**

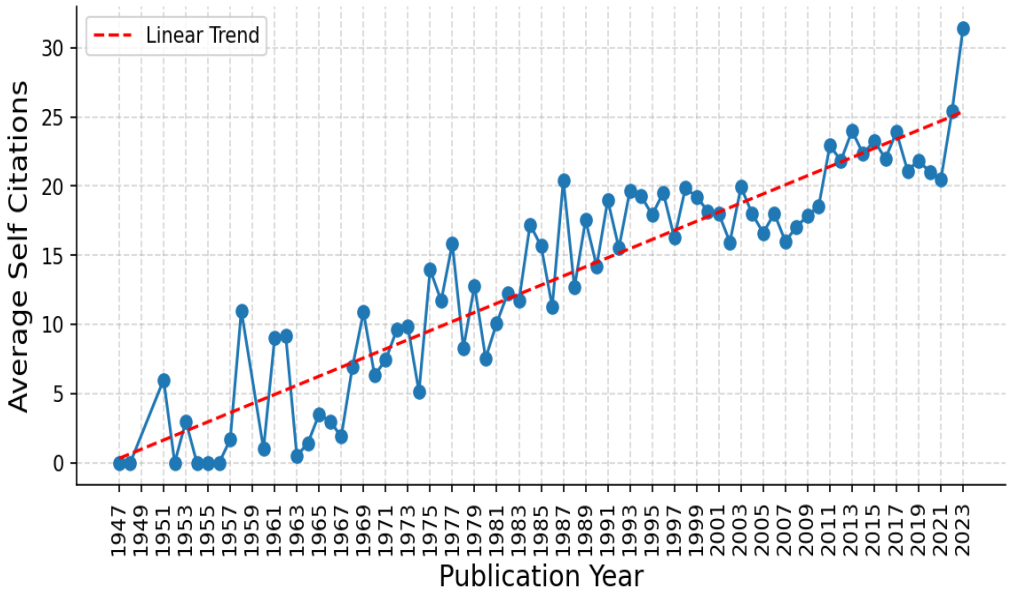


Figure 3. Average self citations over the years. Red dashed line represents the trend line with slope 0.33.

Table 1 provides an analysis of the top 1% highly cited papers affiliated with India, organized by decades. The table includes the count of papers, their percentage contribution to the total, the total citations, share of self citations and whether the papers were published under Open Access (Yes) or not (No). Among the 29,556 highly cited papers, 21.4% (6,327) were published as open access, while 78.59% (23,229) were not open access. The transition to Open Access is evident, with nearly 21% of recent highly cited papers being openly accessible. The production of highly cited papers increased dramatically after 2000, with nearly 90% of the top 1% papers produced between 2001 and 2023. The most significant contribution came from the 2011–2020 decade (49.45%). Citations reflect the growing impact of Indian research globally. Papers from 2011–2020 have received the highest citations (48,98,374), nearly half of the total. This trend demonstrates India's rising contribution to global research impact, particularly in recent decades, through both high citation counts and improved accessibility. In contrast, an analysis of the proportion of self citations to total citations reveals that 6.5% of self citations refer to papers published during 2021–2023, 6.6% to those from 2011–2020 and 5.5% to papers from 2002–2010. Table 2 presents a list of papers in which self citations account for more than 90% of their total citations. These papers are primarily from recent decades and are associated with smaller team sizes.

Table 2. List of papers with more than 90% of self citations.

S. no.	Paper title	Pub year	Team size	Total citations	Self citations	% Self citations
1	Analysis, adaptive control and synchronization of a seven-term novel 3-D chaotic system	2013	2	176	173	98.3
2	Global chaos synchronization of a family of n-scroll hyperchaotic chua circuits using backstepping control with recursive feedback	2013	2	148	145	98.0
3	Sliding controller design of hybrid synchronization of Four-Wing Chaotic systems	2011	2	182	178	97.8
4	Sliding mode control based global chaos control of Liu-Liu-Liu-Su chaotic system	2012	1	165	161	97.6
5	Adaptive anti-synchronization of Uncertain Tigan and Li Systems	2012	2	162	158	97.5
6	Active controller design for generalized projective synchronization of four-scroll chaotic systems	2011	2	164	159	97.0
7	A new eight-term 3-D polynomial chaotic system with three quadratic nonlinearities	2014	1	181	175	96.7
8	Global chaos control of hyperchaotic Liu system via sliding control method	2012	1	163	157	96.3
9	Anti-synchronization of Lu and Pan chaotic systems by adaptive nonlinear control	2011	2	161	155	96.3
10	Global chaos synchronization of hyperchaotic Pang and hyperchaotic Wang systems via adaptive control	2012	2	152	146	96.1
11	A new six-term 3-D chaotic system with an exponential nonlinearity	2013	1	197	189	95.9
12	Generalized Projective Synchronization of Two-Scroll Systems via Adaptive Control	2012	2	162	155	95.7
13	Anti-synchronization of hyperchaotic lorenz and hyperchaotic chen systems by adaptive control	2011	2	158	151	95.6
14	The generalized projective synchronization of hyperchaotic lorenz and hyperchaotic Qi systems via active control	2011	2	165	157	95.2
15	Hybrid synchronization of n-scroll chaotic chua circuits using adaptive backstepping control design with recursive feedback	2013	2	161	152	94.4
16	Analysis, properties and control of an eight-term 3-D chaotic system with an exponential nonlinearity	2015	1	154	145	94.2
17	Generalised projective synchronisation of novel 3-D chaotic systems with an	2014	1	166	155	93.4

	exponential non-linearity via active and adaptive control					
18	Adaptive synchronization of chemical chaotic reactors	2015	1	149	136	91.3
19	Global chaos synchronization of WINDMI and Coullet chaotic systems using adaptive backstepping control design	2014	2	160	146	91.3
20	Analysis, control and synchronisation of a six-term novel chaotic system with three quadratic nonlinearities	2014	1	192	175	91.1
21	Analysis and anti-Synchronization of a novel chaotic system via active and adaptive controllers	2013	1	184	167	90.8
22	Analysis, adaptive control and anti-synchronization of a six-term novel jerk chaotic system with two exponential nonlinearities and its circuit simulation	2015	5	150	136	90.7
23	Analysis and adaptive synchronization of two novel chaotic systems with hyperbolic sinusoidal and cosinusoidal nonlinearity and unknown parameters	2013	1	191	173	90.6

Team size vs. self citations

According to Wuchty et al. (2007), larger teams tend to dominate in producing highly cited research due to their ability to combine various expertise, tackle complex problems, and leverage collaborative networks. In contrast, smaller teams are more likely to focus on niche and innovative topics, which may gain recognition more gradually. This high- lights the role of team size in shaping research impact and citation patterns. Figure 4(a) illustrates the authors with multiple highly cited papers. 73.47% of the authors have a single paper appearing in the top 1% highly cited papers. 13.55% authors have two papers in the highly cited list followed by 5.2% having 3 papers, etc. Figure 4(b) represents the distribution of teams appearing in the paper and the number of publications. A negative correlation exists between the number of papers and the number of authors, suggesting that smaller teams (fewer authors) are more prolific in producing papers. In contrast, larger teams contribute fewer papers, likely due to the increased complexity and coordination involved in collaborative research. As team size grows, the number of papers decreases significantly, with some exceptions for very large teams.

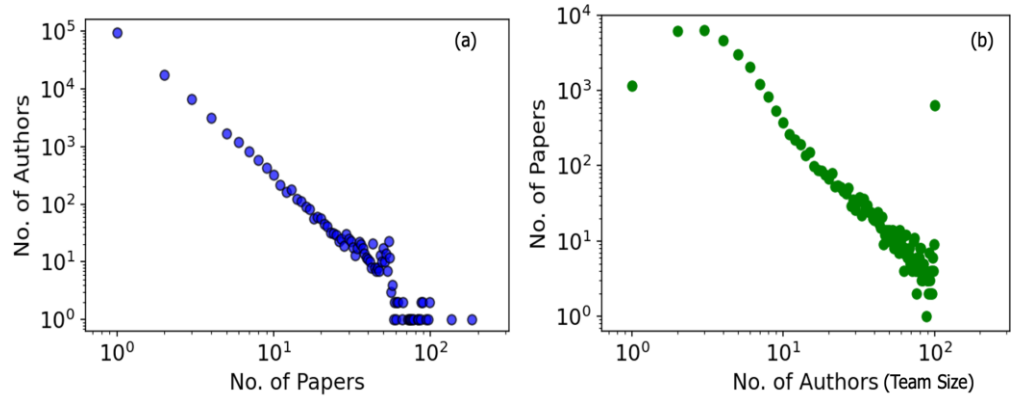


Figure 4. (a) Number of papers vs. count of authors. (b) Team size vs. number of papers.

In addition, Figure 5 represents the average self citations based on team size. The average self citations generally increase as the team size grows, particularly in smaller teams. There are notable peaks and troughs, indicating variability in self citation practices as team sizes change. Teams with fewer members (1–25) exhibit a more consistent, gradual increase in average self citations, indicating relatively steady behavior. The red dashed line illustrates the overall increasing trend, suggesting a positive correlation between team size and average self citations, although the data show variability.

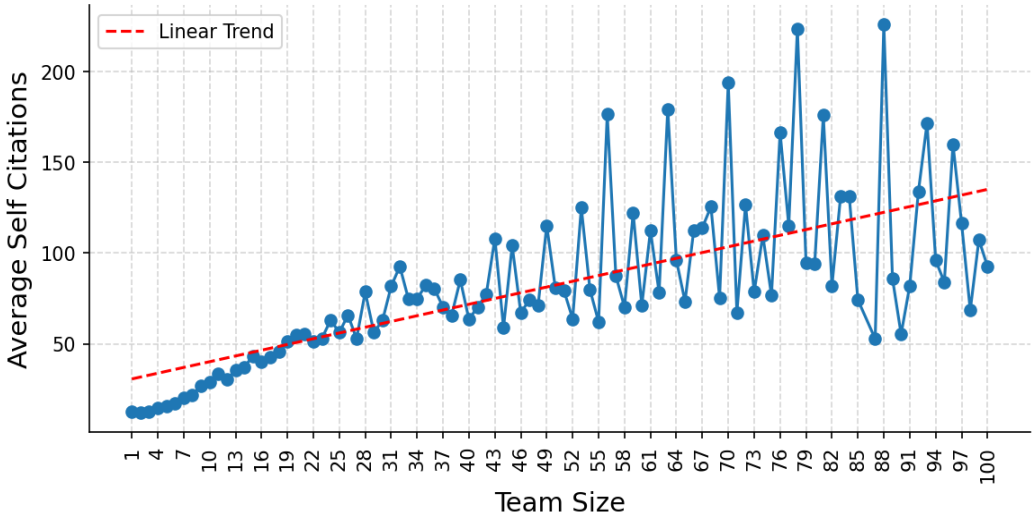


Figure 5. Team size vs average self citations.

Collaboration pattern vs. self citations

An influential and highly cited paper often the result of interdisciplinary teams and collaborative efforts. According to Uzzi et al. (2013), interdisciplinary collaboration and diversity in research teams are significant factors contributing to innovative and impactful research, as they bring together varied perspectives and expertise, which

increase the likelihood of producing groundbreaking work. The table 3 provides an overview of the distribution of highly cited papers and self citations between domestic and international collaborations. Out of the total 29,556 highly cited papers, domestic collaborations contribute the majority, accounting for 56.9%, while international collaborations contribute 43.1%. The total citation distribution indicates that 49.7% of total citations belong to domestic papers, while 50.3% of the total citations belong to international papers. However, when examining self citations, international collaborations dominate with 67% self citations, compared to 33% self citations from domestic collaborations. This indicates that while domestic collaborations produce a higher number of highly cited papers, international collaborations are more likely to generate self citations. This could reflect the broader scope, larger teams, and multidisciplinary nature of international projects, leading to higher interconnectedness and frequent self referencing. In contrast, domestic collaborations, often involving smaller teams, focus on national-level research with relatively fewer self citations.

Table 3. Distribution of number of papers, corresponding total and self citations.

Category	Domestic		International		Total
	Count	In %	Count	In %	
No. of papers	16,823	56.9%	12,733	43.1%	29,556
No. of total citations	48,43,057	49.7%	49,10,563	50.3%	97,53,620
No. of self citations	1,95,752	33%	3,97,569	67%	5,93,321

Figure 6 represents different aspects of the relationship between team size and the nature of collaborations (domes. vs. international). The figure on the left represents the number of publications versus team size where publications domestic collaborations peak at smaller team sizes and gradually decline as team size increases, showing minimal activity for larger teams. International collaborations demonstrate a consistent trend, maintaining a higher level of publications for medium-to-large team sizes compared to domestic collaborations. The spike in international collaborations for the largest team size is likely driven by highly collaborative or global scale projects. Similarly, the right figure represents the number of self citations vs. team size where self citations for domestic collaborations are higher for smaller team sizes and decline sharply as the team size increases. International collaborations exhibit a steadier decline in self citations, with smaller team sizes showing relatively lower self citations compared to domestic ones. There is a noticeable peak for international collaborations at the larger team size (likely an outlier). In general, domestic collaborations dominate smaller team sizes in terms of both self citations and publications, whereas international collaborations gain prominence as team sizes increase. In addition, the USA is the topmost collaborator followed by the UK and Germany.

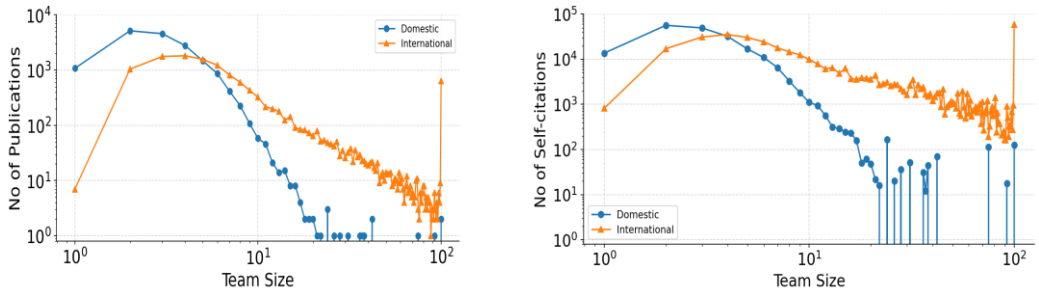


Figure 6. Domestic vs. international collaboration. (left) Team size vs. number of publications. (b) Team size vs. self citations.

Conclusion

In conclusion, citations and self citations play a crucial role in academic research by measuring impact, ensuring research continuity, and fostering collaboration. The study on highly cited Indian research papers highlights that domestic collaborations contribute a greater number of highly cited papers, while international collaborations generate more self citations, likely due to broader networks and multidisciplinary projects. Team size also influences citation patterns, with smaller teams producing more papers but relying more on self citations, whereas larger teams are linked to international collaborations and higher citation visibility. Striking a balance between self-referencing and engaging with the broader academic community is essential to maintain credibility and ensure meaningful research impact.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, citation practices vary across disciplines, making direct comparisons challenging, especially in fields where self citation rates are naturally higher. Second, limitation lies in the classification of team sizes, as it does not account for variations in individual author contributions, which can influence citation impact. Third, the study differentiates between domestic and international collaborations but does not fully capture the complexity of multi-country partnerships that may affect citation trends. Lastly certain fields, such as biomedical research, have higher citation frequencies than others, making direct comparisons across disciplines challenging.

Data availability

The datasets used in the study will be available from the corresponding author on request.

Acknowledgment

Authors gratefully acknowledges the Research and Development Cell, BML Munjal University for their financial support through the seed grant (No: BMU/RDC/SG/2024-06), which made this research possible.

Conflict of interest

The authors declare no conflict of interest.

References

Bornmann, L., Daniel, H.D., 2008. What do citation counts measure? a review of studies on citing behavior. Journal of documentation 64, 45–80.

Costas, R., van Leeuwen, T., Bordons, M., 2010. Self citations at the meso and individual levels: effects of different calculation methods. Scientometrics 82, 517–537.

Fowler, J., Aksnes, D., 2007. Does self citation pay? Scientometrics 72, 427–437.

Hellsten, I., Lambiotte, R., Scharnhorst, A., Ausloos, M., 2007. Self citations, co-authorships and keywords: A new approach to scientists' field mobility? Scientometrics 72, 469–486.

Hillman, J.R., Baydoun, E., 2019. Quality assurance and relevance in academia: a review. Major challenges facing higher education in the Arab world: Quality assurance and relevance, 13–68.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences 102, 16569– 16572.

Hodge, D.R., 2025. Assessing scholarly impact: Conducting bibliometric analyses, in: Handbook of Research Methods in Social Work. Edward Elgar Publishing, pp. 76–83.

Hussein, T.M., Ateeq, A., Ateeq, R.A., Agarwal, S., 2024. Self citation in scholarly work: Balancing self-reference with scientific integrity, in: Business Sustainability with Artificial Intelligence (AI): Challenges and Opportunities: Volume 2. Springer, pp. 361–369.

Hyland, K., 2003. Self citation and self-reference: Credibility and promotion in academic publication. Journal of the American Society for Information Science and technology 54, 251–259.

Martin, B.R., 2013. Whither research integrity? plagiarism, self-plagiarism and coercive citation in an age of research assessment.

Moed, H.F., 2006. Citation analysis in research evaluation. volume 9. Springer Science & Business Media.

Szomszor, M., Pendlebury, D.A., Adams, J., 2020. How much is too much? the difference between research influence and self citation excess. Scientometrics 123, 1119–1147.

Tahamtan, I., Bornmann, L., 2019. What do citation counts measure? an updated review of studies on citations in scientific documents published between 2006 and 2018. Scientometrics 121, 1635–1684.

Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical combinations and scientific impact. Science 342, 468–472.

Van Leeuwen, T., 2013. Bibliometric research evaluations, web of science and the social sciences and humanities: a problematic relationship? Bibliometrie-Praxis und Forschung 2.

Van Noorden, R., Chawla, D.S., 2019. Hundreds of extreme self-citing scientists revealed in new database. Nature 572, 578–580.

Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 316, 1036–1039.

https://doi.org/10.51408/issi2025_011

Shaping Innovation: A Regional Perspective on Industrial PhD Programs in Italy

Tindaro Cicero[1], Annalisa Di Benedetto[2]

[1]*tindaro.cicero@unimercatorum.it*
Department of Engineering and Science, Universitas Mercatorum (Italy)

[2]*annalisa.dibenedetto@istat.it*
Directorate for Social Statistics and Population Census, Italian National Institute of Statistics (ISTAT) (Italy)

Abstract

Doctoral education has evolved into a strategic asset for connecting academic research with industry needs. Industrial PhDs promote collaboration between universities and the private sector, aligning with the Triple Helix model of interaction among academia, industry, and government. In Italy, reforms under the National Research Plan 2015–2020 and Ministerial Decree n. 45/2013 introduced innovative doctoral programs, including industrial and intersectoral PhDs, emphasizing integration with non-academic sectors. These programs benefit from EU FSE/FESR funding, requiring formal agreements with companies, joint project design, and training periods within companies or abroad. In 2021, stricter criteria for Industrial PhDs mandated specific scientific projects and company representation in Steering Committees, enhancing their alignment with industry needs. This study examines Industrial PhD programs in 2022–2023, using text analysis (LDA) on program titles to identify thematic areas like digital transformation, sustainability, and advanced manufacturing. Spatial analysis explores the relationship between program distribution and regional innovation performance.

Preliminary findings suggest a growing alignment of Industrial PhDs with innovation hotspots, as evidenced by changes in spatial distribution and program focus. This indicates strategic diversification, influenced by funding policies and strengthened academia-industry partnerships, fostering innovation and regional economic development.

Introduction

The importance of doctoral education has grown significantly in recent years, not just as a means of advancing academic knowledge but also as a strategic asset in bridging the gap between research and industry needs (Shin et al., 2018). This transformation is particularly evident in the case of industrial PhDs (Roolaht, 2015; Borrell-Damian et al., 2015; Borrell-Damian et al., 2010; Harman, 2008; Thune et al., 2012), which aim to align closely with the needs of modern economies by fostering collaboration between universities and the private sector. The shift in doctoral education reflects broader societal and technological changes, emphasizing practical skills and knowledge transfer relevant to non-academic careers (Bernhard & Olsson, 2020; Haapakorpi, 2017; Jones, 2018).

Industrial PhD programs are increasingly recognized as vital in promoting innovation, particularly within the framework of the Triple Helix model, which highlights the interplay between universities, industry, and government (Thune, 2010; Gustavsson et al., 2016). Doctoral students are increasingly recognized as central to fostering university–industry collaboration, serving as conduits for

knowledge transfer and sharing. Studies have shown that their research activities enhance these interactions, particularly when public policy initiatives actively promote such relationships (Santos et al., 2021; Thune, 2009). By embedding PhD students directly into industry settings, these programs enable a continuous exchange of knowledge and skills that benefit both academic and industrial partners. This approach not only accelerates the application of research but also enhances the employability of graduates in sectors outside academia, addressing the often-cited challenge of underemployment among doctorate holders (Grimm, 2018; Leogrande at al 2022).

Studies have shown that industrial PhD initiatives contribute to regional economic development by leveraging university expertise to solve practical industrial challenges, thereby enhancing competitiveness (Gustavsson et al., 2016). Moreover, they foster a culture of innovation through collaborative projects that bring together diverse stakeholders to co-produce knowledge and technological solutions (Sjöö & Hellström, 2019). This model has been particularly successful in countries like Sweden, where industrial PhDs have been used to strengthen ties between academia and industry, promoting sustainable economic growth (Olsson & Bernhard, 2023).

However, despite these advantages, challenges remain in effectively managing these collaborations. Conflicting priorities between academic and industrial stakeholders can complicate the execution of joint projects, as each party may have different expectations regarding outcomes and timelines (Grimm, 2018; Bienkowska & Klofsten, 2012). Addressing these challenges requires robust frameworks that facilitate communication, trust-building, and mutual commitment, ensuring that industrial PhD programs deliver value to all participants (Bernhard & Olsson, 2020; Thune, 2010).

While much of the existing literature has focused on the broader benefits of industrial PhD programs and their successful implementation in countries like Sweden and Norway (Gustavsson et al., 2016; Thune, 2010; Sjöö & Hellström, 2019), relatively little attention has been paid to their development and characteristics in the Italian context. The dynamics of industrial PhD programs in Italy remain largely underexplored, particularly in terms of their integration within the national doctoral education framework, their relative weight in the overall doctoral system, and their geographical distribution. Existing studies have often focused on local or regional contexts (Compagnucci et al. 2024), thereby limiting a comprehensive understanding of their role at the national level.

This study seeks to fill this gap by providing a detailed analysis of the Italian case, in order to situate Italy's approach within the broader European landascape and allow for an assessment of best practices and policy transferability.

The purpose of this paper is to analyze the characteristics of industrial PhD programs, focusing on their development at the academic level and their geographical distribution. Specifically, the study addresses the following research questions:

i. What are the characteristics of Industrial PhD programs in Italy and in the main European experiences?

ii. What is the proportion of industrial PhDs within the entire set of doctoral programs?

iii. What is their geographical distribution across the territory?
iv. The concentration of industrial PhD programs are linked to the region's innovation performance?
To address these questions, public data provided by ANVUR, the Ministry and the European Commission will be utilized, as described in the specific section. By examining these aspects, this study aims to contribute to the ongoing discussion on how to optimize industrial PhD programs to maximize their impact on innovation, regional development, and the broader goals of economic and social sustainability. Furthermore, this analysis provides a foundation for understanding how the Italian experience compares to international benchmarks, contributing to the global literature on the role of industrial PhDs in fostering innovation and economic development.

The Italian context and the main European experiences

In Italy, with a note dated August 31, 2016, as part of the implementation of the National Research Plan 2015–2020, the Research Ministry introduced significant updates regarding innovative doctorates and work-based learning. The new ministerial guidelines set criteria to differentiate traditional PhDs from innovative doctorates, including the industrial/intersectoral doctorate, which promotes integration with sectors outside academia. These types are not mutually exclusive, with emphasis on valuing combinations among them.
Based on the concept of "collaboration with companies" of the Ministerial Decree n. 45/2013, the ministerial note now clarifies that accredited courses labeled as "Industrial PhDs" can be of two types: (1) courses in partnership with companies, which may reserve positions for employees of one or more companies; (2) conventional doctoral courses that include curricula developed in collaboration with companies. Specifically, the PON aimed at utilizing EU FSE/FESR funds[1] provides co-financing for innovative industrial doctoral courses. These programs require joint design, including for individual PhD students, and offer opportunities for students to spend training periods at companies or abroad.
In 2021, the Italian Ministry of University and Research tightened the criteria for qualifying as an Industrial PhD programme providing specific requirements regarding the collaboration with companies (Ministerial Decree n. 226/2021). It is required for every company involved in an Industrial PhD programme a formal agreement and for each one a company member must be included in the Steering Committee. Furthermore, a specific scientific project must be outlined, consistently with the programme theme and scopes.
Even without the Industrial PhD qualification it is still possible for the programmes to establish other agreements to collaborate with companies that carry out R&D

[1] PON refers to an Italian National Operational Program (PON) that foster economic growth, social cohesion, and regional development utilizing European Union funding sources, specifically the European Social Fund (FSE) and the European Regional Development Fund (FESR). FSE supports projects related to social inclusion, employment, education, and skills development, while FESR finances infrastructure, innovation, and economic development projects, particularly in less-developed regions.

activities without all the requirements above, or even other agreements with any company aiming at scholarship funding. Starting from 2022 these latter kinds of agreements have increased significantly due to the introduction of a co-financing framework within the PNNR (Recovery and Resilience Plan, i.e. the implementation tool, in Italy, for the Next Generation EU program).
This framework produces a three-type classification of the collaborations, as defined in Table 1:

Table 1. Collaboration between PhD programmes and companies' classification.

Collaboration scope	Company requisites	Specificity
Scholarship founding	None	
Associated PhD with companies	Demonstrable coherent and functional R&D activity	- the associated company must finance at least one scholarship for the PhD programme, co-financing can be supported both by the associated company and by external parties (based on specific agreements)
Industrial PhD	- Demonstrable coherent and functional R&D activity, coherent and functional - At least one company member included in the PhD Steering Committee - Outline of a specific scientific project, consistent with the programme * For each company involved	- the company must finance at least one scholarship for the PhD programme, co-financing can be supported both by the company and by external parties (based on specific agreements) - specific requirements can be provided for the research activities (interdisciplinarity, intersectorality) - a portion of the available places for the PhD programme can be reserved to company employees engaged in highly qualified activitieS

Italy's approach to Industrial PhDs aligns with broader European efforts to strengthen academia-industry collaboration. Indeed, since 2011 European Union has included industrial doctorates in its policy agenda for research, innovation and employment. However there are notable differences in implementation across countries, of which have extensive and long-lasting experience in this field. Among the main European experiences, we can certainly mention those of Germany, the UK, France and the Nordic countries, which have developed industrial PhD models with distinct characteristics and diverse approaches to university-business collaboration.
In Germany, Industrial PhD programs function as a collaboration between universities and companies, allowing doctoral students to conduct research while being integrated into an industrial environment. These programs typically involve a contractual agreement where the PhD candidate is employed by a company while

being supervised by both academic and industry mentors. The students divide their time between dissertation research and company-related tasks, gaining hands-on experience in a corporate setting.

In some cases, companies also support PhD students working within university faculties on joint research projects to enhance cooperation between academia and industry. Another common model involves professionals who pursue a PhD while maintaining their regular job in a company, with academic supervision remaining independent. These programs aim to bridge the gap between theoretical research and practical application, fostering knowledge transfer and innovation while equipping students with industry-relevant skills (Grimm, 2018).

In the UK, the main initiative is the Industrial CASE Studentships program, which supports collaboration between academia and industry through industrial PhD opportunities. Established in 1994, the program is administered by UK Research and Innovation (UKRI) and its constituent research councils, such as the Engineering and Physical Sciences Research Council (EPSRC) and the Science and Technology Facilities Council (STFC). The program aims to enhance innovation and equip PhD graduates with skills that meet both academic and industry needs. However, the proportion of PhD scholarships funded by Industrial CASE varies depending on the research council's priorities, available funding, and the level of industry engagement in specific research areas. Under the Industrial CASE scheme, PhD candidates work on projects co-designed by a university and an industrial partner, addressing real-world challenges. The program provides four years of funding, combining academic research with practical industry exposure. The funding includes tuition fee coverage, a stipend (often higher than standard UKRI stipends due to industry contributions), and research costs. Additionally, students are required to spend a minimum of three months working directly with the industrial partner, promoting knowledge transfer and building valuable professional networks. Collaboration models in Industrial CASE include joint knowledge development, applied research to improve products or processes, or exploratory research into emerging technologies[2]. The scheme is distinguished by its integration of academic and practical training, which ensures that PhD graduates are well prepared for careers in both academia and industry, often providing advantages in the private sector (Lee & Miozzo, 2015).The CIFRE (Convention Industrielle de Formation par la Recherche) program in France is a state-supported initiative that fosters collaboration between academia and industry through industrial PhD programs. Managed by the National Association for Research and Technology (ANRT), it has been in place since 1981 with the goal of strengthening university-industry exchanges while enhancing the professional integration of PhD graduates. The program provides three-year funding for PhD candidates employed by companies, requiring a formal agreement between the firm and a public research laboratory. The state grants a scholarship over three years, while the company offers a minimum annual salary. Collaboration models under CIFRE include knowledge transfer from academia to industry, joint knowledge co-

[2]https://www.ukri.org/what-we-do/developing-people-and-skills/stfc/training/types-of-training/industrial-case-studentships/ (last access on April 10th 2025)

development, or outsourcing of research to universities. Research strategies vary from product/process improvement, to developing new scientific competences, or exploring high-risk innovation areas. The program covers approximately 9% of funded PhDs in France and is particularly attractive to firms, including SMEs, as it helps de-risk R&D investments while providing access to academic expertise. Compared to similar European schemes, CIFRE is distinguished by its formalized agreements, structured collaboration, and emphasis on long-term engagement. It is recognized as a key mechanism in bridging scientific research and industrial application, facilitating innovation, and ensuring highly skilled workforce integration into industry (Plantec et al., 2019).

Industrial doctorates in Nordic countries, particularly in Norway and Sweden, have gained prominence as a mechanism to bridge the gap between academia and industry. These programs, often funded or co-hosted by companies, provide doctoral candidates with direct exposure to industrial research environments, fostering collaboration and facilitating smoother transitions into non-academic careers. Unlike traditional PhD paths, industrial doctorates emphasize applied research, aligning doctoral training with industry needs and enhancing employability.

Despite their advantages, the effectiveness of industrial PhDs varies by country. In Sweden, exposure to industry is high, often through structured collaborations, while in Norway, prior industry experience before entering a PhD program plays a more significant role in shaping career transitions. However, the transition to industry is not always automatic, as skill mismatches persist, requiring graduates to actively build networks during their PhD. While university-industry partnerships provide opportunities, personal networking remains crucial in securing industry positions.

Ultimately, industrial doctorates contribute to bridging academia and industry, yet their success depends on the strength of university-industry ties and the ability of doctoral candidates to leverage these connections. The Nordic model highlights the potential of structured collaborations but also underscores the need for stronger institutional support in facilitating career transitions. (Germain-Alamartine et al., 2021).At the European level, in 2014, the Marie Skłodowska-Curie Actions (MSCA) introduced the Industrial Doctoral Programmes flagship initiative, designed to foster PhD training through partnerships between universities, companies, and other socio-economic stakeholders. Funded by the European Union's Horizon Europe program, MSCA provides substantial financial support, covering salaries, research costs, and mobility allowances. This funding directly supports both the PhD candidate and the host organizations, incentivizing international collaboration. The MSCA initiative stands out for its global outlook, interdisciplinary scope, and comprehensive support for mobility and training, setting it apart from more localized and industry-specific national industrial PhD programs. A central feature of MSCA programs is the emphasis on international, intersectoral, and interdisciplinary mobility. PhD candidates are required to work in multiple countries and often across academia and industry, fostering global collaboration. Consequently, MSCA programs are

particularly attractive for building international networks and preparing PhD graduates for global careers in both academia and industry[3].
All Industrial PhD schemes typically involve shared funding between an agency, a university, and a company, with the company applying for the grant. The PhD student, enrolled in a regular program, is jointly supervised by both institutions and splits their time between the university and the firm (Thune & Børing, 2015).
Italy has developed a structured system for industrial PhDs, overall its regulatory approach is more prescriptive compared to the company-driven models in Germany or the flexible, incentive-based schemes in France and the UK. The introduction of PNRR funds has significantly expanded industry-academia collaborations, but challenges remain in ensuring long-term private sector engagement beyond co-financing mechanisms.2.

Data and methodology

This explorative research on Italian Industrial PhD programs was conducted over a two-year period (2022-2023) and all the scientific disciplines. Since 2013, the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR) has been entrusted with conducting the initial accreditation and annual verification of PhD programmes. ANVUR therefore verifies that PhD programs meet specific requirements, the evaluation procedure is mainly based on a set of ex-ante indicators focused on the quality of the PhD Steering Committee and of the Scientific Coordinator, the teaching activities, the financial sustainability, the availability of scholarships, the research infrastructures, and the overall coherence of the research project.
Table 2 underscores the regional disparities in the availability and uptake of doctoral education in Italy, with larger and more populous regions generally hosting more extensive programs and enrollments. It presents in detail the distribution of PhD programs and students at the NUTS 2 level across Italian regions for the academic years 2022 (XXXVIII cycle) and 2023 (XXXIV cycle). It highlights both regional and national trends in higher education, reflecting the heterogeneity of doctoral education in Italy. The number of the accredited PhD programs and students for 2022 (XXXVIII cycle) and 2023 (XXXIX cycle) is reported in Table 2. The data was derived from the public website of ANVUR[4] and from the Portal of Higher Education Data of the Italian Ministry of University and Research[5].

[3] https://marie-sklodowska-curie-actions.ec.europa.eu/actions/doctoral-networks (last access on April 10th 2025)
[4] www.anvur.it (last access on November 09th, 2024)
[5] https://ustat.mur.gov.it/ (last access on November 09th, 2024)

Table 2. Number of universities, PhD programs and PhD students in 2022 and 2023, by NUTS 2 level.

NUTS 2 level	Number of universities	2022 (XXXVIII cycle)		2023 (XXXIV cycle)	
		PhD programs	PhD Students	PhD programs	PhD Students
Piedmont	4	56	1024	56	1039
Aosta Valley	1	0	0	0	0
Liguria	1	30	480	31	491
Lombardy	15	162	2790	171	2937
Abruzzo	5	38	393	41	402
Molise	1	7	62	7	56
Campania	10	102	1490	113	1785
Apulia	5	57	841	64	777
Basilicata	1	5	82	5	40
Calabria	4	27	288	31	243
Sicily	4	70	768	71	931
Sardinia	2	25	276	28	214
Autonomous Province of Bolzano/Bozen	1	8	105	7	65
Autonomous Province of Trento	1	18	290	18	333
Veneto	4	72	1029	74	1246
Friuli Venezia Giulia	3	36	389	37	393
Emilia-Romagna	4	98	2552	101	1508
Tuscany	8	96	1310	106	1354
Umbria	2	23	202	25	224
Marche	4	28	383	29	380
Lazio	20	194	2495	204	2458
Italy	**100**	**1152**	**17249**	**1219**	**16876**

In 2022, there were 1,152 PhD programs nationwide with 17,249 enrolled students, while in 2023, these numbers shifted slightly to 1,219 courses and 16,876 students. This indicates an increase in the number of PhD programs but a slight decrease in overall student enrollment. The regional distribution shows notable differences: Lombardy, with its 15 universities, leads in both years, offering 162 courses to 2,790 students in 2022 and increasing to 171 courses for 2,937 students in 2023. Lazio follows with 20 universities, offering 194 courses with 2,495 students in 2022 and 204 courses with 2,458 students in 2023. Both regions account for a significant portion of Italy's doctoral education system.

In contrast, smaller regions like Aosta Valley, Molise, and Basilicata have minimal or no representation in doctoral education, with Valle d'Aosta reporting no PhD programs or students in either year. Regions like Emilia Romagna and Tuscany also demonstrate strong participation, with substantial numbers of courses and students, though Emilia Romagna shows a marked decline in student enrollment from 2,552 in 2022 to 1,508 in 2023, despite a slight increase in courses offered.

Southern regions such as Campania and Apulia show a growing number of courses but varying trends in student enrollment, with Campania experiencing a significant rise in students from 1,490 in 2022 to 1,785 in 2023, while Puglia sees a reduction from 841 to 777. Sicily, on the other hand, reflects consistent growth, increasing both courses and student numbers between the two years.

The data also emphasizes the contributions of autonomous provinces like Autonomous Provinces of Trento and Bolzano, which, despite their smaller size, maintain a consistent presence in doctoral education. Trento, for instance, reported stable course offerings at 18 but increased student enrollment from 290 in 2022 to 333 in 2023.

The subsequent section of this study will delve into the mapping of PhD programs specifically characterized as industrial doctorates for the academic years 2022 and 2023. To analyze these programs, text analysis techniques, particularly Latent Dirichlet Allocation (LDA), will be applied to the titles of the industrial PhD programs. This methodology will allow for the identification of key thematic areas addressed by these programs, shedding light on the specific industrial and technological challenges they aim to tackle. By clustering and categorizing topics, this analysis will highlight trends, such as the prevalence of themes related to digital transformation, sustainability, or advanced manufacturing, providing a clearer understanding of the strategic focus of these doctoral initiatives.

In addition, potential relationships between the geographical distribution of industrial PhDs and specific territorial characteristics will be explored thought spatial descriptive statistics. The main aim is to investigate whether the presence and concentration of industrial PhD programs are linked to the region's innovation performance.

Results and discussion

Geographic distribution of industrial PhD programs

Table 3 provides an overview of the distribution of industrial PhD programs in Italy at the NUTS 2 level for the XXXVIII (2022) and XXXIV (2023) cycles, highlighting an overall growth both in absolute and relative terms. The total number of industrial PhD programs increased from 49 (4.3% of the total PhD programs) in 2022 to 83 (6.8%) in 2023. This growth reflects an expanding emphasis on the alignment between doctoral education and industrial needs, in line with broader European trends promoting university-industry collaboration (Etzkowitz & Leydesdorff, 2000). Regionally, the data reveal significant disparities. Liguria shows a remarkable increase in industrial PhD programs, rising from 23.3% to 41.9% of the total PhDs in the region, positioning it as a leader in integrating doctoral training with industrial applications. Similarly, regions such as Abruzzo (18.4% to 22%) and Umbria (21.7% to 36%) demonstrate significant relative growth, reflecting targeted regional initiatives. Conversely, several regions, including Basilicata, Calabria, Sardinia, and Veneto, report no industrial PhD programs, underscoring persistent challenges in fostering such programs in less industrialized or peripheral areas.

The data also underline the prominence of certain industrial and academic hubs, such as Lombardia and Lazio, which exhibit modest relative growth but play critical roles due to their overall academic and industrial capacity. Notably, Molise shows a decline in the relative share of industrial PhD programs (from 57.1% to 28.6%), which may warrant further investigation into the underlying causes. The increasing proportion of industrial PhD programs at the national level (from 4.3% to 6.8%)

signals a growing recognition of their strategic importance for enhancing research and innovation ecosystems.

Table 3. Number and percentage of industrial PhD programs in 2022 and 2023, by NUTS 2 level.

NUTS 2 level	Number of Industrial PhD programs (percentage in brackets)	
	2022 (XXXVIII cycle)	2023 (XXXIV cycle)
Piedmont	1 (1.8)	1 (1.8)
Aosta Valley	0 (0)	0 (0)
Liguria	7 (23.3)	13 (41.9)
Lombardy	1 (0.6)	5 (2.9)
Abruzzo	7 (18.4)	9 (22)
Molise	4 (57.1)	2 (28.6)
Campania	9 (8.8)	10 (8.8)
Apulia	1 (1.8)	2 (3.1)
Basilicata	0 (0)	0 (0)
Calabria	0 (0)	0 (0)
Sicily	3 (4.3)	8 (11.3)
Sardinia	0 (0)	0 (0)
Autonomous Province of Bolzano/Bozen	0 (0)	0 (0)
Autonomous Province of Trento	1 (5.6)	1 (5.6)
Veneto	0 (0)	0 (0)
Friuli Venezia Giulia	1 (2.8)	5 (13.5)
Emilia-Romagna	1 (1)	5 (5)
Tuscany	2 (2.1)	3 (2.8)
Umbria	5 (21.7)	9 (36)
Marche	0 (0)	0 (0)
Lazio	6 (3.1)	10 (4.9)
Italy	49 (4.3)	83 (6.8)

The maps reported in Figure 1 complement the data presented in Table 3 by offering a geographic visualization of the distribution of industrial PhD programs across Italian regions at the NUTS 2 level for the XXXVIII (2022) and XXXIV (2023) cycles. They highlight the persistence of significant regional disparities in the adoption of industrial PhDs, with marked differences between northern, central, and southern Italy.

In 2022, central regions such as Umbria and Abruzzo emerged as leaders in industrial PhD adoption, while northern and southern regions generally showed lower percentages. In 2023, Liguria demonstrated a notable increase, positioning itself alongside Umbria as a leader in integrating industrial PhD programs. However, several southern regions, including Basilicata, Calabria, and Sardinia, remain largely excluded from this trend, reflecting ongoing challenges in fostering university-industry collaboration in less industrialized or peripheral areas.

The maps visually emphasize the growing polarization, with industrial PhD programs concentrating in specific academic and industrial hubs. This uneven geographic distribution highlights the need for targeted policies to support

underrepresented regions, enabling broader national alignment with the European agenda for university-industry collaboration.

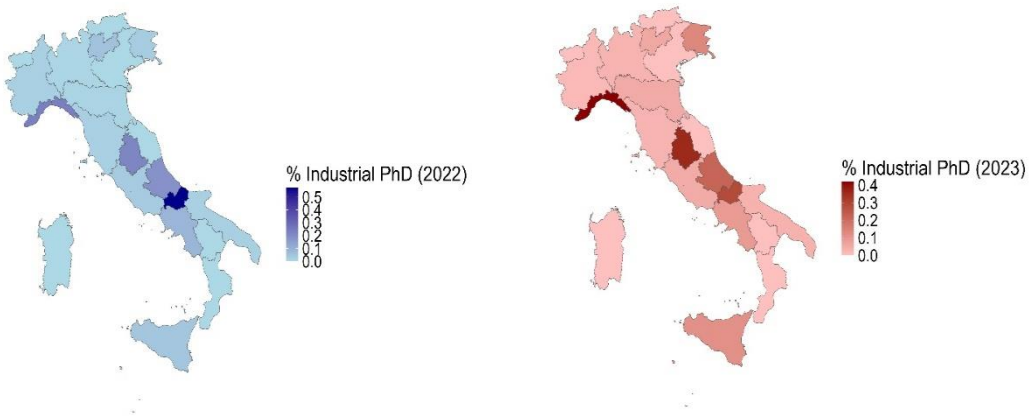


Figure 1. Percentage of 2022 and 2023 Industrial PhD program.

Thematic distribution of Industrial PhD programs

Figure 2 illustrates the application of Latent Dirichlet Allocation (LDA) to identify the main terms extracted from the titles of industrial PhD programs in Italy for the years 2022 and 2023. The analysis was performed on filtered datasets containing only industrial PhD programs, and the titles were translated into English to ensure consistency.

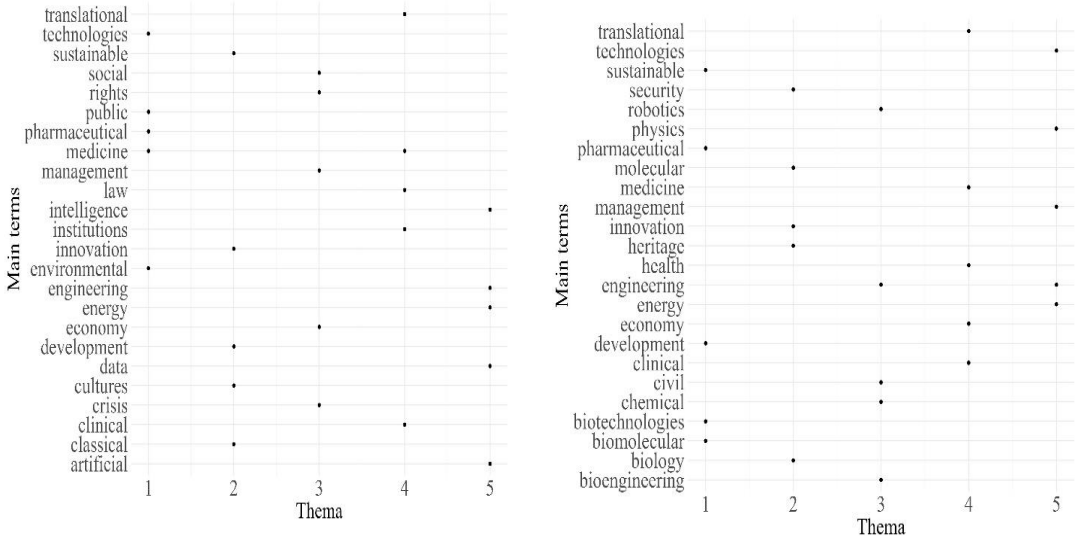


Figure 2. LDA analysis for 2022 and 2023 Industrial PhD programs.

In 2022, the thematic distribution highlights a strong emphasis on interdisciplinarity and sustainability, with terms such as "translational", "sustainable", "engineering", and "management" prominently represented. The presence of keywords such as "medicine", "pharmaceutical", and "environmental" suggests that healthcare, pharmaceutical research, and environmental studies played a significant role in shaping industrial PhD offerings. Furthermore, the inclusion of terms like "social" and "rights" points to an integration of social sciences, complementing the technical and scientific focus. The distribution of terms across the identified themes reflects a diverse approach to doctoral education, addressing a broad spectrum of societal and industrial challenges.

In 2023, the thematic landscape demonstrates a notable evolution, with an increasing emphasis on advanced technologies and specialized scientific domains. Terms such as "robotics", "physics", "biotechnologies", and "bioengineering" emerge as key elements, reflecting a shift towards cutting-edge fields with strong industrial applications. Despite this shift, the prominence of terms like "sustainable" and "innovation" indicates the continued prioritization of sustainability and the alignment of doctoral education with contemporary global challenges. Additionally, the emergence of terms such as "heritage" and "civil" suggests a growing recognition of cultural and infrastructural dimensions within industrial PhD programs.

A comparison of the two years reveals a dynamic evolution in the focus areas of industrial PhD programs in Italy. While the 2022 programs exhibit a broader thematic distribution, encompassing healthcare, sustainability, and social sciences, the 2023 programs signal a more targeted orientation towards technology-driven and specialized research fields. This shift underscores the responsiveness of doctoral education to emerging trends and evolving industry needs, reflecting the increasing integration of advanced technologies and interdisciplinary approaches. The consistent presence of sustainability and innovation as core themes highlights the strategic role of industrial PhD programs in fostering research and innovation ecosystems that address both industrial priorities and societal challenges.

Statistical analysis

The spatial distribution of 2022 and 2023 Industrial PhD programs can be effectively represented through Standard Deviational Ellipses. The Standard Deviational Ellipse is a graphical representation that shows the orientation, shape, and spatial dispersion of a set of points, its centre corresponds to the centroid (or barycentre) of the spatial distribution (for a more in-depth and technical disclosure see Wong & Lee, 2005; Brunsdon & Comber 2015). This representation can incorporate the weight of a variable by adjusting the size and orientation of the Ellipse based on the variance and distribution of that variable (i.e. the number of Industrial PhD programs), allowing it to reflect not only the spatial arrangement of points but also the intensity or significance of specific factors that influence the distribution. In this case, since all distributions consider the spatial centroids of the Italian regions as the set of points, the observable differences in the ellipses and the barycentre can be attributed merely to the weight of the variables considered.

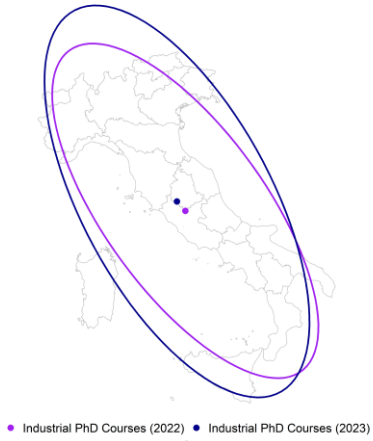


Figure 3. Standard Deviational Ellipses of 2022 and 2023 Industrial PhD programs.

The comparison of the Industrial PhD programs' Standard Deviational Ellipses for 2022 and 2023 (Figure 3) reinforces the key trend already highlighted in the discussion of Figure 1. Moreover, several noteworthy insights emerge when comparing the distribution of Industrial PhD programs with that of all PhD programs and universities.

While the 2022 ellipse for industrial PhD programs is notably narrower than that of all doctoral programs and universities, reflecting a higher concentration of industrial PhDs in a limited number of key hubs, the 2023 ellipse shows a significant shift. In 2023, the ellipse becomes more similar in size and orientation to those of the broader doctoral programs and university locations.

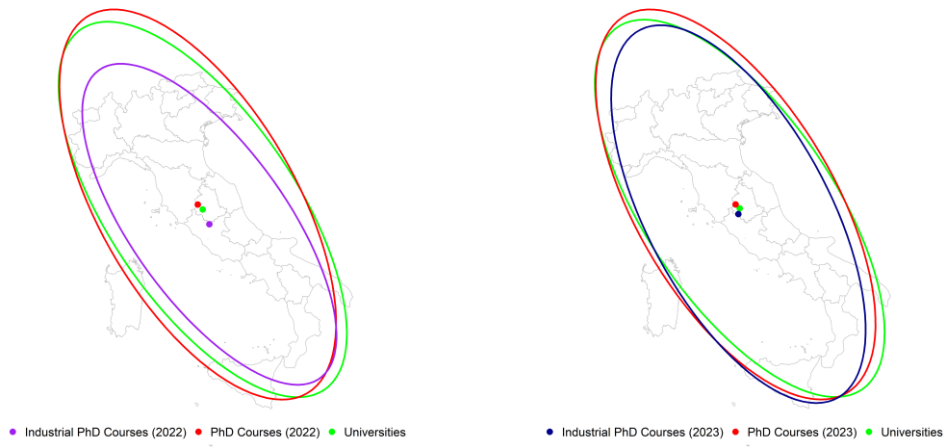


Figure 4. Standard Deviational Ellipses of 2022 and 2023 Universities, PhD programs and Industrial PhD programs.

A simple explanation could be found in the longer time elapsed for the for the XXXIV (2023) cycle since the formalization of the criteria for qualifying an

industrial PhD, revised by the relevant Italian Ministry in 2021 (thus maybe with short notice for the XXXIII cycle). Nevertheless, this change could also be attributed to a growing diversification in the institutions offering industrial PhD programs, possibly driven by policy initiatives aimed at fostering this kind of programs or the increased adoption of collaborative research models across a wider array of universities. Additionally, the expansion may reflect the alignment of local academic and industrial ecosystems with national and European funding priorities, which increasingly emphasize inclusive and distributed research excellence. As a result, the spatial footprint of industrial PhDs appears to be converging with the broader academic landscape, suggesting a gradual diffusion of opportunities beyond the traditional innovation hubs.

An analysis of industrial PhDs cannot ignore the characteristics of the educational and production systems in which they are embedded. For this purpose, the Regional Innovation Scoreboard (RIS) represents a shared and consolidated framework to characterize the territories at the NUTS 2 level in terms of innovation performance, enabling a comparative perspective and addressing various aspects of utmost importance for this study.

The Regional Innovation Scoreboard (RIS) is a report published by the European Commission since 2009 to evaluate the innovation performance of European regions (complementing the European Innovation Scoreboard (EIS), which focuses on national performance). It aims to identify regional differences in innovation capabilities and highlight best practices; it is therefore a particularly fitting reference for this analysis.

The RIS provides a solid set of innovation indicators (including R&D investment, patents, entrepreneurial activities, and education) and following the same methodology of the EIS classifies the European's regions into four Innovation Performance groups according to their Regional Innovation Index (RII[6]):

1. Innovation Leaders (regions with above-average performance);
2. Strong Innovators (regions performing close to the EU average);
3. Moderate Innovators (below-average performers);
4. Emerging Innovators (lowest-performing regions).

[6] RII is calculated as the unweighted average of the normalised scores of 21 indicators. Since RIS uses fewer indicators (21 compared to 32 in the EIS), some with different definitions, and regional data are less timely than the country level data, it is necessary to align the country level results between RIS and AIS. The following correction is therefore applied to the composite indicator scores:

1) Calculate the ratios of the EIS 2023 Summary Innovation Index at country level with that of the EU: EIS_index_CTR / EIS_index_EU;

2) Calculate the ratios of the RIS 2023 Regional Innovation Index at country level with that of the EU: RIS_index_CTR / RIS_index_EU;

3) Calculate the correction factor by dividing the ratios 1) and 2).

These country correction factors are then multiplied with the RII for each region in the corresponding country to obtain final RII scores. Then relative performance scores are calculated by dividing the RII of the region by that of the EU and multiplying by 100. For trend performance, RIIs for all years are divided by that of the EU in 2016 (see the Regional Innovation Scoreboard 2023 – Methodology Report).

Italy is a Moderate Innovator within the EIS, but regional performance differences are high. Referring to the 2021 performance 12 of the 20 Italian regions were Moderate Innovators, but there were also seven Strong Innovators (see Figure 5) and two Emerging Innovators (Calabria and Aosta Valley).
Interestingly in 2023 RIS only three regions still result as Strong Innovators (Emilia-Romagna, Friuli-Venezia Giulia and Autonomous Province of Trento), and two as Emerging Innovators (Sicily and Sardinia), but RII indicator compares the regional performance to that of the EU in the same year. It is also noticeable that 2023 RIS highlights that Italian region performance has increased at a higher rate than that of the EU for all regions compared to 2014, and most strongly for Marche and Abruzzo (Figure 5).

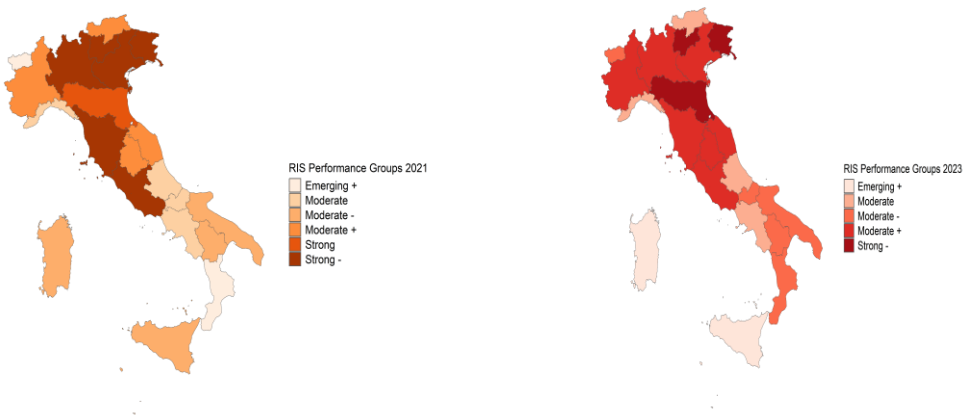


Figure 5. Innovation Performance groups (RIS 2021 and 2023).

In the following analysis, will be therefore considered the performance groups reported in the RIS 2021 and 2023 documents. However, to ensure data comparability, the detailed value of the Regional Innovation Index 2021 and 2022 presented in the 2023 RIS report will be used (see Figure 6).
Since the accreditation procedures for PhD programs occur during the academic year preceding their start, the RII 2021 values will be considered relevant for the XXXVIII cycle (starting in 2022), while the RII 2022 values will be considered relevant for the XXXIV cycle (starting in 2023).

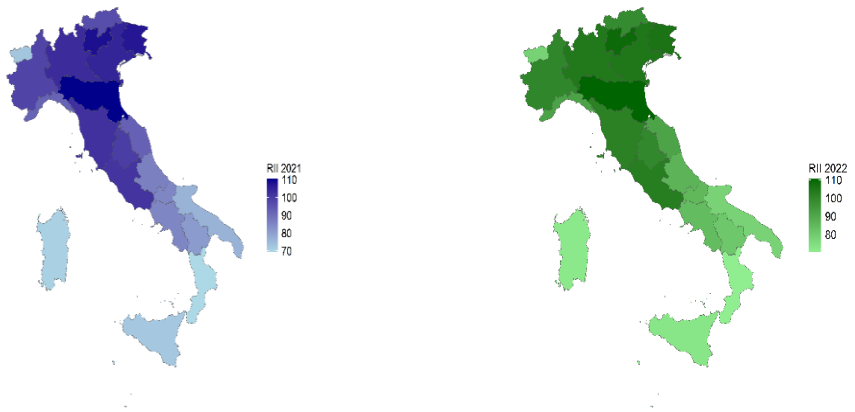


Figure 6. Regional Innovation Index – RII 2021 and RII 2022 (RIS 2023).

The Standard Deviational Ellipses of the RII indicators for 2021 and 2022 are extremely close (almost overlapping, as seen in Figure 7), as are their centroids. It is only possible to highlight a slight widening of the ellipse in the northwest direction between the two years examined (it is important to highlight that much of the data used in the calculation of the RII is not updated annually, and as a result, the indicator inherently exhibits a certain level of stability between updates).

The comparison with the ellipses of the distribution of industrial PhD programs in 2022 and 2023 offers more points of interest. The first is that the centroids of the industrial PhD distributions (for both years under consideration) are located further south than those of the RII.

The second is that the ellipse of the industrial PhD distribution for 2023 has a width and directional orientation more similar to that of the innovative performance distribution, in line with what was expected based on the hypotheses outlined above.

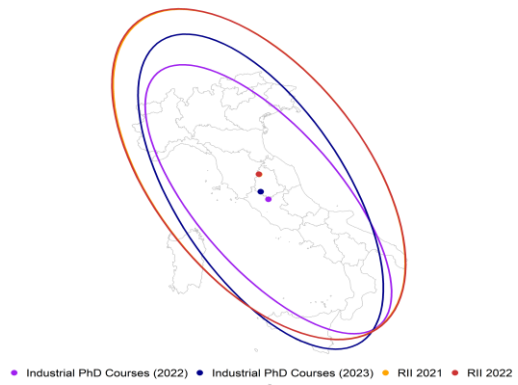


Figure 7. Regional Innovation Index – RII 2021 and RII 2022 (RIS 2023).

The fact that the centroids of the industrial PhD distributions for both 2022 and 2023 are located further south than those of the RII might suggest that industrial PhD programs are becoming more concentrated in southern regions. This could reflect a deliberate policy shift or a regional emphasis on developing innovation and industry-academic collaborations in areas traditionally less involved in these activities. It could also indicate a broader trend of industrial PhDs expanding outside the established innovation hubs, potentially due to regional development initiatives or universities seeking to align with national priorities.
The observation that the ellipse for the industrial PhD distribution in 2023 has a width and directional orientation more like that of the innovative performance distribution may indicate a closer alignment between industrial PhDs and the broader innovation landscape. This could suggest that the 2023 cohort of industrial PhDs is increasingly influenced by or integrated into areas of high innovative activity. Such a change in the shape and direction of the ellipse may also imply that industrial PhD programs are diversifying geographically and aligning more closely with regions that show stronger innovation performance, possibly driven by new funding policies or more strategic collaborations between universities and industries in these areas.

Conclusions

In a recent systematic literature review, Compagnucci and Spigarelli observed that research interest in industrial PhDs has grown rapidly since 2015, attributing this trend to both the emergence of the Third Mission of universities and policy factors, particularly in Europe. Also in the context of this growing attention Compagnucci and Spigarelli's analysis highlighted the marginality of quantitative studies, and pointed out the need for more structured studies, particularly those with a longitudinal perspective to estimate the impact of these types of programs.
This study represents a simple starting point for analysis in the Italian context; nevertheless, it highlights the evolving role of Industrial PhD programs in Italy as strategic tools for fostering collaboration between academia and industry and contributing to regional innovation dynamics. Besides, Italy represents a particularly interesting context, especially considering its regulatory model, more rigid and top-down compared to the industry-led approach in Germany or the more flexible, incentive-driven systems in France.
The analysis reveals significant findings related to the geographic and thematic distribution of these programs.
Between 2022 and 2023, the number of Industrial PhDs increased substantially, with their share rising from 4.3% to 6.8% of all doctoral programs. Regions like Liguria, Umbria, and Abruzzo demonstrated notable growth in industrial PhDs, while southern regions such as Basilicata, Calabria, and Sardinia lagged behind, underscoring persistent disparities. Thematic analysis using Latent Dirichlet Allocation (LDA) identified a shift in focus from broad themes like sustainability and interdisciplinarity in 2022 to more specialized domains such as robotics, biotechnologies, and advanced manufacturing in 2023.
The spatial distribution analysis suggests a growing alignment between Industrial PhD programs and regions with higher innovation performances, although the

centroids of their distribution are located further south compared to those of regional innovation indicators. This finding may indicate a deliberate policy shift to promote innovation in less-developed areas or an emerging trend of universities and industries in southern regions increasing their engagement in collaborative research. However, the persistence of regional disparities calls for broader policies to ensure equitable access to these programs and their benefits.

Future research should aim to evaluate the long-term impact of Industrial PhD programs on regional economic growth, workforce development, and the competitiveness of innovation ecosystems. In particular, the role of Italy's National Recovery and Resilience Plan (PNRR) in shaping the distribution, thematic focus, and effectiveness of these programs requires further exploration. The PNRR provides a unique opportunity to strengthen academic-industry collaboration through co-financed scholarships and investments in innovation-driven education. Assessing the extent to which these resources address regional and national priorities will be crucial to understanding their broader impact.

From a policy perspective, it is essential to address regional imbalances by introducing targeted funding mechanisms for less-developed areas, incentivizing companies to engage in collaborative research, and supporting universities in building capacity for industrial partnerships. Additionally, fostering interdisciplinary approaches and integrating sustainability into the design of Industrial PhDs will be critical to addressing complex societal and industrial challenges. Policymakers should also prioritize the development of robust performance monitoring frameworks to measure the effectiveness of these programs in delivering tangible benefits, including innovation outputs, economic development, and improved employability of graduates. By aligning national and regional policy goals with the strategic objectives of Industrial PhDs, Italy can maximize the potential of these programs as a cornerstone of its innovation and education policy framework, contributing to sustainable and inclusive economic growth. From an European perspective, a more coordinated initiative in the field of industrial PhDs would be highly desirable. Such an effort could help harmonize national systems, facilitate cross-border mobility of doctoral candidates, and promote shared standards for industry-academia collaboration. It would also support the development of a more integrated innovation ecosystem across the EU, strengthening the competitiveness of European research and industry in the global landscape.

References

Bernhard, I., & Olsson, A. K. (2020). University-industry collaboration in higher education: Exploring the informing flows framework in industrial PhD education. Informing Science, 23, 147.

Bienkowska, D., & Klofsten, M. (2012). Creating entrepreneurial networks: academic entrepreneurship, mobility and collaboration during PhD education. Higher Education, 64, 207-222.

Borrell-Damian, L., T. Brown, A. Dearing, J. Font, S. Hagen, J. Metcalfe, and J. Smith. 2010. "Collaborative Doctoral Education: University-Industry Partnerships for Enhancing Knowledge Exchange." Higher Education Policy 23 (4): 493–514. https://doi.org/10.1057/hep.2010.20.

Borrell-Damian, L., R. Morais, and J. H. Smith. 2015. Collaborative Doctoral Education in Europe: Research Partnerships and Employability for Researchers Report on Doc-Careers II Project. Brussels: European University Association.

Brunsdon, C., & Comber, L. (2015). "An Introduction to R for Spatial Analysis and Mapping." SAGE Publications.

Compagnucci, L., & Spigarelli, F. (2024). Industrial doctorates: a systematic literature review and future research agenda. Studies in Higher Education, 1–28.

Compagnucci, L., Spigarelli, F., Perugini, F., & Iacobucci, D. (2024). Industrial doctorates for regional development: the case of Le Marche Region. *Higher Education*, 1-21.

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. Research policy, 29(2), 109-123.

Germain-Alamartine, E., Ahoba-Sam, R., Moghadam-Saman, S., & Evers, G. (2021). Doctoral graduates' transition to industry: networks as a mechanism? Cases from Norway, Sweden and the UK. Studies in Higher Education, 46(12), 2680-2695.

Grimm, K. (2018). Assessing the Industrial PhD: Stakeholder Insights. Journal of Technology and Science Education, 8(4), 214-230.

Gustavsson, L., Nuur, C., & Söderlind, J. (2016). An impact analysis of regional industry—University interactions: The case of industrial PhD schools. Industry and Higher Education, 30(1), 41-51.

Haapakorpi, A. (2017). Doctorate holders outside the academy in Finland: Academic engagement and industry-specific competence. Journal of education and work, 30(1), 53-68. https://doi.org/10.1080/13639080.2015.1119257

Harman, K. M. 2008. "Challenging Traditional Research Training Culture: Industry-Oriented Doctoral Programs in Australian Cooperative Research Centres." In In Cultural Perspectives on Higher Education, edited by J. Välimaa and O. H. Ylijoki, 179–195. Dordrecht: Springer.

Lee, H. F., & Miozzo, M. (2015). How does working on university–industry collaborative projects affect science and engineering doctorates' careers? Evidence from a UK research-based university. The Journal of Technology Transfer, 40, 293-317.Leogrande, A., Costantiello, A., & Laureti, L. (2022). The Impact of New Doctorate Graduates on Innovation Systems in Europe. Available at SSRN 4209643.

Olsson, A.K. & Bernhard, I. (2023). Transforming doctoral education: Exploring industrial PhD collaboration in Sweden. International Journal of Work-Integrated Learning, 24(4), 523-536. https://www.ijwil.org/files/IJWIL_24_4_523_536.pdf

Plantec, Q., Cabanes, B., Le Masson, P., & Weil, B. (2019, June). Exploring practices in university-industry collaborations: the case of collaborative doctoral program in France. In R&D Management 2019.

Roolaht, T. (2015). Enhancing the industrial PhD programme as a policy tool for university—industry cooperation. Industry and Higher Education, 29(4), 257-269.

Santos, P., Veloso, L., & Urze, P. (2021). Students matter: The role of doctoral students in university–industry collaborations. Higher Education Research & Development, 40(7), 1530-1545.

Shin, J. C., Kehm, B. M., & Jones, G. A. (2018). The increasing importance, growth, and evolution of doctoral education. *Doctoral education for the knowledge society: Convergence or divergence in national approaches?*, 1-10.

Sjöö, K., & Hellström, T. (2019). University–industry collaboration: A literature review and synthesis. Industry and higher education, 33(4), 275-285.

Thune, T. (2009). Doctoral students on the university–industry interface: a review of the literature. Higher Education, 58, 637-651.
Thune, T. (2010). The training of "triple helix workers"? Doctoral students in university–industry–government collaborations. Minerva, 48, 463-483.
Thune, T., & Børing, P. (2015). Industry PhD schemes: Developing innovation competencies in firms?. Journal of the Knowledge Economy, 6, 385-401.
Thune, T., S. Kyvik, S. Sörlin, T. B. Olsen, A. Vabø, and C. Tømte. 2012. PhD Education in a Knowledge Society: An Evaluation of PhD Education in Norway. Nordic Institute for Studies in Innovation, Research and Education. Report 25/2012. ISBN 978-82-7218-846-6.
Wong, D. W., & Lee, J. (2005). Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS. John Wiley & Sons.

https://doi.org/10.51408/issi2025_013

Social Impact Analysis of Retracted Paper in the Context of Public Health Emergencies

Liu Xiaojuan[1], Shen Jianing[2], Dai Xinran[3], Yu Yao[4]

[1]*lxj_2007@bnu.edu.cn*, [2]*15639172472@163.com*,
[3]*daixinran0258@163.com*, [4]*yuyao990824@163.com*
School of Government, Beijing Normal University, No. 19, Xinjiekouwai Street,
Haidian Beijing (China)

Abstract

Analyzing the impact of COVID-19 retracted papers can provide references for effectively preventing and controlling negative effects. In this study, 253 COVID-19 retracted papers in Retraction Watch were selected as research objects. Focusing on the paper publication and the retraction notice release, this study analyzes their social impact from three aspects: social attention, public dissemination and policy making. Meanwhile, this study takes typical retracted papers as examples to analyze the impact cascade phenomenon it may trigger. The results show that paper characteristics, delay in retraction, and reasons for retraction play an important role in the social impact of COVID-19 retracted papers, which is highly concentrated. The faster papers gains public attention, the longer the duration of their attention will be. Some papers could be used in policy documents soon after publication, often by referring to the conclusions and discussion sections to enhance persuasion. On this basis, this study proposes strategies to prevent and control the negative impact of retracted papers. First, journals should pay attention to the standardization of the retraction process and statement. Second, researchers should consider public needs and emphasize the social value of scientific research. Third, the supervision department should play an important role in accelerating the process of academic purification through news media and social media. When utilizing academic achievements, policymakers should adequately assess the quality of papers and update retraction information promptly.

Introduction

Retraction serves as a self - correction mechanism within the scientific community, aiming to purify and uphold scientific research ethics. In recent years, there has been an increase in the number of retracted papers due to data issues, image issues, authorship issues, plagiarism and false reviews. The number of retracted papers worldwide per year has risen from 41 in 2000 to over 10,000 in 2023 (Van Noorden, 2023), hitting an all-time high. Retracted papers may confuse subsequent research

with erroneous data or opinions, even mislead practice or decision-making in the wider society. This harms human health, public safety or social development.

Public health emergency is a major infectious disease outbreak or mass unexplained disease that occurs suddenly and may cause serious damage to public health. To effectively prevent, control and eliminate its harm, China formulated *the Regulations on Public Health Emergencies* in 2003. It emphasizes that medical, monitoring, scientific research and other institutions should obey the unified command of the headquarters and concentrate on relevant scientific research work. Academic achievements are disseminated and utilized in academic circles and all sectors of society, providing decision-making support regarding public health emergencies. After the outbreak of COVID-19, international medical journals have responded to the severe situation caused by the epidemic from three aspects: speeding up peer review, open access, and improving data mining and analysis tools (Shen, 2022). The rapid publication of a large number of academic papers has not only greatly facilitated scholarly communication and information sharing, vaccine research and clinical practice in the field of COVID-19, but has also attracted widespread attention from the government and the public, playing an important role in the formulation of epidemic prevention and control policies (Ren et al., 2023; Ren & Yang, 2023), the analysis of the "infodemic" phenomenon (Geng, 2020), and responses to it (Caulfield, 2020; Li et al., 2021). However, there were also very serious retraction problems during that period (Yeo-Teh & Tang, 2022), involving many top medical journals such as *the New England Journal of Medicine* and *the Lancet*. For example, *The Lancet* published a paper reporting that the use of hydroxychloroquine was associated with a higher risk of ventricular arrhythmias and increased in-hospital mortality among COVID-19 patients (Mehra et al., 2021). The results of this study led some countries to ban the use of hydroxychloroquine for the treatment of COVID-19 and suspend clinical trials. This study was later retracted due to uncertain data authenticity. The World Health Organization then restarted trials of the drug hydroxychloroquine. When the paper was published, it attracted great attention worldwide and was reported by 236 mainstream media outlets on the same day. It was mentioned more than 5,000 times on Twitter, blogs, and other social media platforms. Ultimately, it not only shook the public perception but also had a significant negative impact on clinical practice. This suggests that the social impact of retracted papers is not static. Rather, it evolves at landmark events such as paper publication, multiple challenge investigations, and retraction notice releases. When the epidemic prevention and control entered a stable period, scholars conducted in-depth research on COVID-19. Academic papers published in the early stage of the

epidemic were retracted and even a series of retractions resulted from large-scale investigations. The "positive impact" of some papers before the retraction may hide major errors, which are potentially harmful and should not be ignored. Therefore, it is necessary to explore the social impact triggered by retracted papers, especially focusing on the impacts of these papers on public cognition and policy-making. This will provide references for effective prevention and control of the negative effects.

Since the concept of retraction was first introduced in the 1980s, academics have begun to focus on several aspects: the construction of the retraction system (Resnik et al., 2015; Yang, 2020), the basic characteristics of retracted papers (e.g., time, subject, and country distribution) (Song & Yang, 2023), the characteristics of retractions (e.g., reason for retraction, delay in retraction) (Rubbo et al., 2022; Sun et al., 2023), as well as the academic impact (Yuan & Jin, 2024) and social impact (Khan et al., 2022; Liu, Wang, et al., 2022) after retraction. Focusing on the field of COVID-19, the social impact of retracted papers is mostly reflected in the Altmetric Attention Score (AAS). Khan et al. (2022) found that the 22 retracted papers in their study received a great deal of attention in social media, with Twitter and Mendeley being the most popular media platforms. However, the datasets of the existing studies are mostly limited to the period before 2021 and have not yet covered the data during the stable period of the epidemic. This leads to a smaller amount of valid data for the study, which may affect the comprehensiveness and accuracy of the conclusions. Existing studies mainly focus on static analysis of AAS, lacking a dynamic perspective to offer in-depth interpretations of the data and a thorough understanding of its development and evolution. Additionally, these studies focus on the distribution characteristics of altmetric indicators, but overlook detailed content analysis.

The study focuses on COVID-19, using Retraction Watch and Altmetric.com as the main data sources. Combining the landmark events in the life cycle of a retracted paper, it explores the social impact and negative effects of these actively or passively "disappeared" retracted papers. The several research questions are proposed as follows:

Q1: How do retracted papers acquire attention in the social field from a dynamic perspective?

Q2: What social impact do retracted papers generate across social attention, public dissemination, and policy making during a public health emergency, and what potential negative effects may they trigger? Especially in terms of policy making, what are the motivations for mentioning retracted papers in policy documents, and what content are mentioned?

Q3: Is it possible for a retracted paper to trigger a cascading impact in both

academic and social fields? how this impact unfolds, and what consequences it leads to?

Data collection

This study searched for papers with "COVID-19" or "SARS-CoV-2" in the title from Retraction Watch database. In addition, the papers listed under "Retracted coronavirus (COVID-19) papers" were also included in the dataset. A total of 328 records were obtained, involving 299 retracted papers with basic information, reason for retraction, and retraction time. The data collection was completed on June 28th, 2023. To ensure the accessibility of the subsequent data, conference abstracts, conference papers, and preprints were excluded. Ultimately, this study obtained 253 papers, whose publication and retraction time are shown in Figure 1. Altmetrics data collection was completed on August 13, 2023, including AAS, values of altmetric indicators, etc., for retracted papers.

In terms of discipline distribution, the 253 COVID-19 retracted papers cover all major disciplines of Retraction Watch (shown in Table 1). Health science was the most predominant, followed by business and technology. Since 106 papers belong to more than one discipline, double counting was carried out in this study. Pharmacology had the highest number of retracted papers among all sub-disciplines. Two hundred and fifty-three papers were from 65 countries and regions. Fifty-nine of these papers were multinational collaborations, and only the country of the first author was counted. The United States and China tied for first place, both with 40 retracted papers, representing 15.8% of the total 253 retracted papers; the Republic of Malta came in third. Of the 28 retracted papers, 27 were from the same journal, *Early Human Development*. Twenty-one of these papers were retracted on the same day, but none of the retraction notices mentioned a specific reason for the retraction. India, Pakistan, Spain, the UK, Egypt, Brazil, and Iran rank from 4th to 10th.

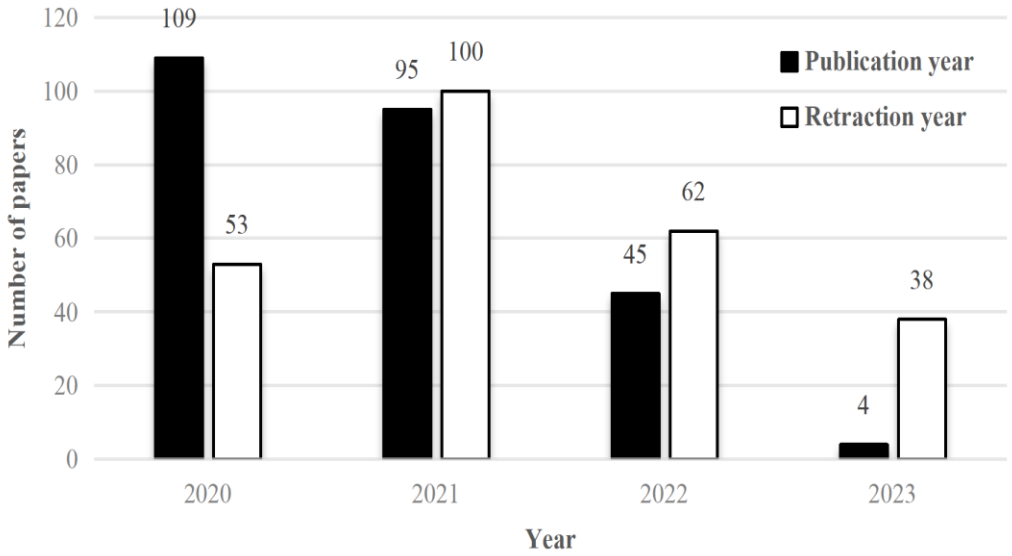


Figure 1. Publication and time distribution of 253 COVID-19 retracted papers.

Table 1. Discipline distribution of COVID-19 retracted papers.

Discipline	Sub-discipline (number of retracted papers)	Number of retracted papers
Health Sciences	Pharmacology (245), Public Health and Safety (51), Biostatistics/Epidemiology (38), Occupational Health and Safety (19), Radiology and Imaging (7), etc.	250
Business and Technology	Business (19), Data Science (14), Technology (14), Computer Science (12), etc.	55
Social Sciences	Sociology (14), Education (13), Psychology (9), Communication (2), etc.	41
Basic Life Sciences	Toxicology (7), Microbiology (7), Biochemistry (7), etc.	23
Physical Sciences	Physics (2), Geology (1), etc.	5
Environmental Sciences	Environmental Science (4), Food Science (1), etc.	5
Humanities	Journalism (3), etc.	5

Analysis method

Serious retraction problems erupted during the COVID-19 pandemic, damaging the scientific research ecosystem jointly built by researchers, the public, and the government. It aroused the concern of the academic community about the value of academic achievements. Meanwhile, the community is also paying close attention to the impact of retracted papers on policy-making, public opinion and the social environment. To gain a deeper understanding of the impact of retracted papers among

a wider audience, this study focuses on two landmark events: paper publication and retraction notice release. On the one hand, once a paper is published, its social impact will follow and the paper's characteristics may affect the public's attention and cognition. On the other hand, the release of retraction notices marks the change of the paper from normal to retraction, and the impact of the retracted paper may change. As the key indicators of the landmark event, the delay in retraction and reason for retraction could provide an important reference basis for analyzing the potential social impact of retracted papers. To explore the entire process of purification of scientific research environment, we characterize social impacts with the help of altmetric indicators. On this basis, this study will systematically analyze the multidimensional impacts generated by COVID-19 retracted papers. The research framework is shown in Figure 2.

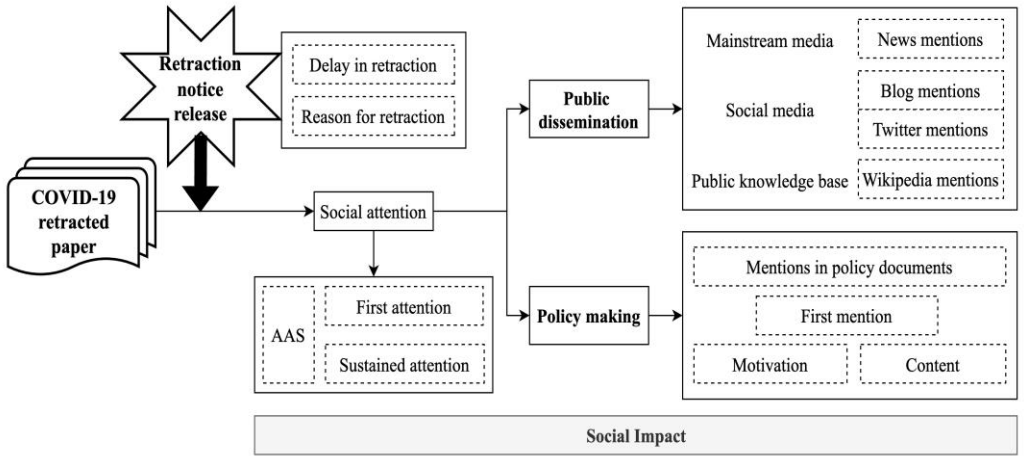


Figure 2. Research framework.

Retraction notice

This study analyzes retraction notices from two aspects: delay in retraction and reason for retraction. The delay in retraction is the time interval between paper publication and the release of retraction notice, which is an important indicator to measure the purification timeliness of retracted papers. Rapid response and timely action by academic institutions or journals can curb the potential negative impact of retracted papers. This study found that 35 of the 253 retracted papers were retracted on the day of publication. Of these, 22 papers did not specify the reason for retraction, and 7 papers were duplicates due to publisher error. The mean delay in retraction for the remaining 218 retracted papers was 249.5 days, with a median of 175 days. The article with the longest delay in retraction is *A topic-based hierarchical*

publish/subscribe messaging middleware for COVID-19 detection in X-ray image and its metadata. After 952 days of publication, this article was retracted along with other articles for academic misconduct, including false peer review and improper citation, in a series of retractions from Soft Computing on May 29, 2023.

Due to the complexity of the reasons for retraction, a unified classification system has not yet been formed. Referring to existing research, the study analyzes 253 paper retraction notices and classifies the reasons for retraction into academic misconduct (137 papers, accounting for 54.2%) (Bar-Ilan & Halevi, 2018) and scientific error (66 papers, accounting for 26.1%) (Ma et al., 2023; Xie et al., 2022). In addition, some papers (58 papers, accounting for 22.9%) were classified as "other" due to the absence of a retraction statement or lack of a specified reason for retraction. The specific distribution is shown in Table 2. Among them, 67 papers involved multiple reasons for retraction and were counted repeatedly. Academic misconduct usually includes plagiarism, inappropriate authorship, ethical violations, and so on. Scientific error is more concerned with problems in scientific research in terms of the rigor of experimental design, reliability of data sources, and accuracy of methodology, including incorrect/unreliable data, incorrect/unreliable results, and so on. Retracted papers containing scientific distortion and unreliable knowledge are considered as a barrier to the advancement of science (Bar-Ilan & Halevi, 2018). Especially in public health emergencies, academic achievements play an indispensable role in supporting epidemic prevention and control. Therefore, the impact of retracted papers due to scientific errors is particularly crucial.

Table 2. Distribution of reasons for retraction of COVID-19 retracted papers.

Primary classification of reasons for retraction	Secondary classification of reasons for retraction	Number of retracted papers	Proportion
Academic Misconduct	False peer review	42	16.94%
	Duplicate publication due to publisher error	22	8.87%
	Improper citation	21	8.47%
	Violation of experimental ethics	19	7.66%
	Duplicate publication	18	7.26%
	Plagiarism	16	6.45%
	Inappropriate attribution	13	5.24%
	Conflict of interest	7	2.82%

	No data rights	5	2.02%
	Copyright notice	4	1.61%
	Artificial Intelligence Generated Content	4	1.61%
Scientific Error	Incorrect/unreliable results	45	18.15%
	Incorrect/unreliable data	26	10.48%
	Analysis error	20	8.06%
	Text error	7	2.82%
	Method error	6	2.42%
	Image error	4	1.61%
Other	No specific reason for retraction	49	19.76%
	No retraction notice	10	4.03%

Social impact

Social impact refers to the influence or benefit that academic achievements bring to public cognition, public policy, public service, economy or culture. Many scholars use Altmetrics as a potential indicator for measuring social impact, effectively supplementing traditional scientometrics with diverse and comprehensive data sources (F. Guo et al., 2016; L. Guo & Zhou, 2023). Among them, AAS can reflect the degree of attention to paper outside the academic community. Yu Houqiang et al. (2014) divide altmetrics indicators into three levels of dissemination, access and utilization to analyze the deepening degree of the social impact. González-Betancor S M et al. (2023) consider that each type of digital platform where a paper is mentioned reflects a different dimension of influence than the academic one: media influence (mentions in mainstream news), social media influence (mentions in Twitter), educational impact (mentions in Wikipedia) and political influence (mentions in public policy reports). It is possible to quantify the task of knowledge transfer to society multidimensionally (Arroyo-Machado et al., 2022). Combined with the rich data provided by Altmetric.com, this study explores the impact of retracted papers through public dissemination and policy making.

At the social attention level, the first attention marks the moment the paper first gains public prominence. The interval between paper publication and first attention reflects the timeliness of the paper's social attention. The sustained attention is the time interval between the last AAS update and the first attention. Considering that some papers still receive attention during data collection, the sustained attention of these

papers is set as the interval between the first attention and data collection date (2023/08/13). For public dissemination, the social impact of COVID-19 retracted papers is spread through diversified media. This study measures the dissemination through different media, including news reports (mainstream media), blog mentions (social media), Twitter mentions (social media), and Wikipedia mentions (public knowledge bases). At the policy making level, the number of policy documents mentioning the papers is used to measure the paper's utilization. The motivation and method of mentioning the paper in the policy documents are analyzed to further understand the interactive relationship between academic research and policy-making, which reflects the impact of academic achievements on policy-making.

Results

Social attention

AAS is calculated based on the attention of various sectors of society, such as government departments, mainstream news media, social networking sites and peer review platforms. It is usually considered to reflect the social impact produced by the papers. Seventy-nine of 253 retracted papers had an AAS of 0, accounting for 31.2%. The remaining 174 papers had an average AAS of 1,044.67. The 8% of the retracted papers (designated as Papers A) accounted for 85% of the AAS. This suggests that the social impact of retracted papers is highly concentrated. The majority of Papers A belong to the field of health sciences, involving topics such as therapeutic drugs, comorbidity studies, vaccination and mask protection effects. This differs from the main topics of highly cited papers (citation frequency >100), which are more related to therapeutic drugs, complication research and public mental health. In Papers A, the United States ranks first where the papers are from, accounting for 50%, and there is only one paper from China. The reasons for retraction are mostly in the category of scientific errors, including erroneous/unreliable results, data, and analysis, which reflects a more concentrated and heightened public concern for scientific errors. In the context of global public health emergencies, scientific knowledge is crucial for policy-making and public health. Once scientific errors in retracted papers are revealed, they may undermine the public trust in academia, and may even interfere with the formulation and effective implementation of relevant prevention and control policies. A similar phenomenon can be observed in a wider dataset. For example, Serghiou S et al. (2021) collected retracted papers from 2010 to 2015 across multiple disciplines. They found that the main reason for retractions of the most popular papers with an AAS of >20 was that the research results were unreliable.

The first attention and sustained attention reveal the timeliness and ongoing interest of social attention aroused by COVID-19 retracted papers. On average, 174 papers received social attention for the first time in 44.5 days after publication. Twenty-six papers (accounting for 14.9%) aroused social attention and discussion on the same day of publication. The average duration of attention was 298.6 days, with a maximum of 1,238 days. There were 50 papers (accounting for 28.7%) whose attention lasted only 1 day, which was a flash in the pan and was quickly overwhelmed by other information. In this study, the intervals between the first attention of the paper and its publication were counted using time intervals of 1 day, 7 days, 30 days, and 180 days. The distribution of specific intervals and their average duration of attention are shown in Figure 3. As the first attention interval decreased, the duration of attention increased significantly, indicating that the paper was able to gain social attention in a shorter period. Even after active or passive retractions, due to the popular topic labels, the negative impacts generated by papers cannot be effectively controlled immediately, and continue to trigger discussions over a longer period.

To explore how the rapidly generated social impacts of retracted papers change and their possible negative effects, this study takes *Facemasks in the COVID-19 era: a health hypothesis* (Vainshelboim, 2021) published in *Medical Hypotheses* and *6-month consequences of COVID-19 in patients discharged from hospitals: a cohort study* (Huang et al., 2021) published in *The Lancet* as examples. The two papers were published around the same time, and both quickly attracted social attention on the day of publication. They both lasted for more than 900 days and had AAS of more than 10,000. However, the evolution of their social impact is different. Specifically, the former did not cause significant social repercussions at the early stage of publication. However, on April 10, 2021, the authors of the paper posted a tweet related to the paper, which was deleted by the Twitter platform later that day. This series of events quickly triggered an outburst of attention from social media users, with 21,855 tweets in 11 days, producing a huge social impact. As the third-party agency issued a statement and the journal editorial board launched an investigation, the paper was formally retracted due to improper authorship, improper citation, and unreliable data. Since then, its social attention has declined, and the delay in retraction was 162 days. After the retraction notice release, a portion of the public still referred to the paper to support their personal views. Therefore, the potential negative effects of the paper persisted. The latter gained high social attention on the day of publication. The mainstream media, as the main force of dissemination, reported the paper more than 110 times on that day. As the author published

subsequent related research results, the paper continued to receive attention from society and was mentioned several times in policy documents. It was not until a reader questioned the data in November 2022 that the journal immediately launched an investigation and issued a notice of concern. Ultimately, the paper was officially retracted and republished with a statement six months later for data errors. The delay in retraction of the paper was 882 days, spanning multiple critical stages of the outbreak. The social impact and potential negative effects of the retraction cannot be ignored. As the paper was quickly republished after being retracted and the topic involved faded in popularity, the number of mentions of the original retracted paper on major platforms dropped significantly.

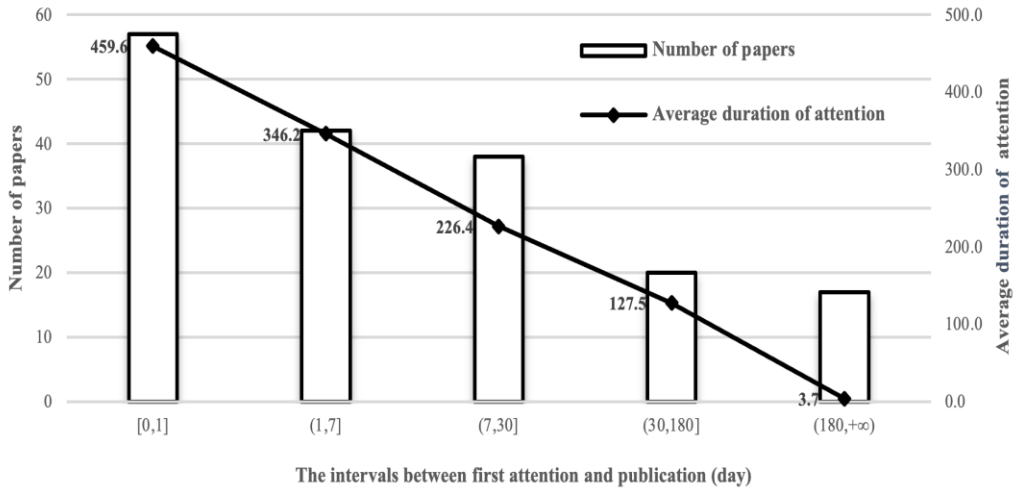


Figure 3. First attention distribution and average duration of attention.

Public dissemination

This study uses the numerical values and coverage of typical indicators to measure the dissemination of papers, as shown in Table 3. The coverage rate of altmetrics indicators, such as mainstream media and social media, of COVID-19 retracted papers is more than 40%. Some papers are mentioned multiple times by Wikipedia. It indicates that these papers' dissemination platforms are diverse and their social impact is wide. Twitter has the highest coverage of mentions, with a mean value of 1,608.95, and the overall dissemination intensity is relatively high. The main dissemination channels of COVID-19 retracted papers are consistent with the existing research (Liu, Sun, et al., 2022). The number of mentions on Twitter for different papers varies widely, with a range reaching up to 45,584. In contrast, blogs and mainstream media have a more balanced dissemination. Notably, retracted

papers that were widely reported by over 50 news media tended to have a shorter delay in retraction. Most of them were retracted within 50 days after publication, which could be a potential positive effect of media attention on the timeliness of purification. In addition, Wikipedia, as a public knowledge base with the core values of openness, inclusiveness and collaborative sharing, plays an important role in the dissemination and popularization of knowledge. Eighteen retracted papers were mentioned 113 times by Wikipedia entries. These entries cover (1) the terminology associated with COVID-19 and its complications, including the therapeutic agents like azithromycin, ivermectin and hydroxychloroquine; (2) the retraction records of academic achievements and instances of academic misconduct; and (3) the latest progress of related clinical trial programs. These entries provide the public with a wealth of professional, authoritative and continuously updated information to meet their concerns and requirements on the global issue of COVID-19. However, Wikipedia's public collaborative editing mechanism is unable to synchronize and update retractions in entries promptly, which contributes to the retention and continued dissemination of misleading information on the platform to some extent.

Table 3. Value of typical altmetric indicators.

Indicator	Coverage ratio	Mean	Median	Standard deviation	Maximum	Minimum
News Mentions	42.5%（74）	75.26	8	239.96	1 692	1
Blog Mentions	52.9%（92）	5.23	1	14.14	107	1
Twitter Mentions	92.5%（161）	1 608.95	5	5 467.36	45 585	1
Wikipedia Mentions	10.3%（18）	6.27	2.5	9.16	33	1

Note: The value in brackets of coverage ratio is “the number of papers with non-zero indicator values”.

Policy making

The mention of academic achievements in policy documents is an important manifestation of their social impact. Especially in public health emergencies, academic papers provide important scientific guidance for relevant policy-making, which has led to an increased emphasis on science in policy decisions (Ren et al., 2023; Yin et al., 2021). Fourteen COVID-19 retracted papers were mentioned in 41 policy documents, among which 10 papers were all retracted due to scientific errors, including incorrect/unreliable data or results, and analysis errors. To further explore the possible negative effects of COVID-19 retracted papers on the scientificity of

epidemic prevention policies, the study analyzed the characteristics of policy document mentions from three aspects: first mention, motivation of mention, and content of mention. Excluding two policy documents for which the original text was not available, 39 policy documents were obtained as a sample. Referring to existing research (Yu et al., 2023) and combining the experience in the coding process, the study developed a content analysis coding table for the motivations and content mentioned in policy documents, as shown in Tables 4 and 5.

Table 4. The coding table of motivation mentions in policy documents.

Primary coding	Secondary coding	Explanation
M1 Background Mention		Introduce an issue and explain the background or significance of the policy.
M2 Support Mention	M2.1 Source Support	Provide sources for concepts, data, theories in policy documents.
	M2.2 Methodological Support	Justify the research methodology or data processing of the policy document.
	M2.3 Argument support	Provide support for arguments, including conclusions or facts.
M3 Construction Mention	M3.1 Indicative orientation	Indicate relevant papers to rich background knowledge or trace the origins of different research perspectives.
	M3.2 Argument base	Formulate new ideas based on the content of the paper.
	M3.3 Meta-analysis	Meta-analyze the data, models from papers as the research content of policy documents.
	M3.4 Scientific review	Review scientifically, discuss and even criticize the papers mentioned.
M4 Unable to Judge	M4.1 Incidental mentions	Mention in appendices, reports or papers included in the policy documents.
	M4.2 Pure mention	No element of the paper is mentioned, and the motivation is vague.

Table 5. The coding table of content mentions in policy documents.

Primary coding	Secondary coding	Explanation
C1 Content	C1.1 Title	Mention the title of the paper exactly.
	C1.2 Abstract	Mention the abstract of the paper, either completely or partially.
	C1.3 Methodology	Mention the methods or models applied in the paper.
	C1.4 Conclusion	Mention the conclusions, discussion, and recommendations of the paper.
C2 Entity	C2.1 Fragments	Mention fragments of the paper, including concepts, ideas, diagrams, paragraphs, etc.
	C2.2 Tools	Mention software, websites, databases, etc., used in the paper.
C3 Generalization	C3.1 Topics	Mention the topic, central question, or research area of the paper.
	C3.2 Overview	Briefly describe, summarize, or evaluate the paper's main content.
	C3.3 Indirect mention	Mention multiple papers in a single sentence and summarize their common features in a particular aspect.
C4 Pure links		No element of the paper's content is mentioned.

This study quantifies the speed of academic achievements in influencing policy documents by the interval between publication and the first mention. The study found that papers were first mentioned in policy documents on average 77.5 days after their publication. Eleven retracted papers (78.6%) were mentioned in policy documents within 180 days of publication. This indicates that the COVID-19 epidemic has strengthened collaboration and dialogue between academics and policymakers, thereby expediting the translation of academic knowledge into policy-making. In addition, this phenomenon is also related to the shortened release cycle of epidemic prevention policies. For example, *the COVID-19 Clinical Management: Dynamic Guidelines* issued by the World Health Organization is updated at least twice a year, ensuring that the recommendations and standards are always based on the latest scientific evidence. However, high-intensity dialogue between the two parties may lead to an inadequate assessment of the quality of papers in policy documents. Due to the controversial nature and unreliable knowledge of retracted papers, the degree of effect on policy development needs to be further assessed.

In terms of motivation, policymakers introduce academic achievements into policy documents, aiming to promote the transformation of knowledge from academic research to policy decision-making, and improve the scientific nature of policies, and enhance the pertinence and implementation effect of policies. The analysis results of the motivation are shown in Table 6. It was found that 64.1% were to find relevant

evidence for policy documents. Research related to pathological manifestations, complications and antiviral drugs based on COVID-19 can support the development of more effective preventive measures, especially as the arguments and data in policy documents. For example, *Clinical Management of COVID-19: Living Guideline* issued by the World Health Organization states that "there is no research that demonstrates a significant effect of antihypertensive medications on the patient's clinical course, and it is generally recommended to continue using such medications." This argument is supported by the paper *Cardiovascular disease, drug therapy, and mortality in COVID-19*. However, after the article was retracted on June 4, 2020, a series of dynamic guidance documents issued from January 25, 2021, to January 13, 2023, continued to mention the paper as the evidence and did not mark its retraction status. The second most common type is "Construction mention", using elements such as data and models as the foundation for viewpoints in policy documents. A small number of policy documents were designed to analyze the risk of bias in the papers. For example, the *COVID-19 Rapid Guideline: Managing COVID-19* conducted a scientific review of the paper *Remdesivir efficacy in COVID-19 treatment: a randomized controlled trial*. The NICE Expert Advisory Group was seriously concerned about the risk of bias. This paper was retracted 191 days after the guideline was released. In addition, there was one policy document that referred to papers in the appendix section, stating only when and why they were retracted, but without mentioning the motivation.

In terms of mentioned content, this study categorizes the mentions based on the structure of papers to understand which parts of papers have had a significant impact on policy formulation. The results are shown in Table 6. In the dataset, 46.2% of the policy documents mentioned the contents of the conclusion and discussion, which corresponds to the "argument support" with the highest proportion of motivation. Secondly, there is a high proportion of "overview" and "indirect reference". The former mostly summarizes the main content of the paper or evaluates the possible risk of deviation. The latter does not directly mention the specific content of a paper, but summarizes the common features of several papers in one sentence, making the reference source richer. The study concludes that policy documents are more focused on the research content. When mentioning content, policymakers tend to choose conclusions that have practical support for the policy document itself.

Table 6. The coding results of motivation and content mentions in policy documents.

Mention of motivation coding	Proportion	Mention of content coding	Proportion
M1 Background mention	5.1%	C1 Content	51.3%
M2 Support mentions	64.1%	C1.1 Title	0
M2.1 Source support	17.9%	C1.2 Abstract	5.1%
M2.2 Methodological support	0	C1.3 Methods	0
M2.3 Argument support	46.2%	C1.4 Conclusion	46.2%
M3 Construction mention	25.6%	C2 Entity	2.6%
M3.1 Indicative orientation	12.8%	C2.1 Fragments	2.6%
M3.2 Argument base	0	C2.2 Tools	0
M3.3 Meta-analysis	7.7%	C3 Generalization	41.0%
M3.4 Scientific review	5.1%	C3.1 Topics	2.6%
M4 Unable to judge	5.1%	C3.2 Overview	17.9%
M4.1 Incidental mention	2.6%	C3.3 Indirect mention	20.5%
M4.2 Pure mention	2.6%	C4 Pure links	5.1%

Impact cascade phenomenon of retracted papers

Cascade refers to the chain reaction in which an event or behavior triggers a series of related events or behaviors. In the process of citation diffusion, a paper triggers a series of subsequent citations, which is called a citation cascade. Similarly, in the process of information diffusion, the information spreads layer by layer among social media users, forming a huge cascade. The above two are intertwined, which together constitute the impact of COVID-19 retracted papers in academia and society, like ripples spreading on the surface of the water, triggering a sustained and extensive chain effect.

To demonstrate more intuitively the possible negative effects of retracted papers in academia and society, this study selected the paper with the highest AAS as a typical case. The title of the paper is *Ivermectin for prevention and treatment of COVID-19 infection: a systematic review, meta-analysis, and trial sequential analysis to inform clinical guidelines* (noted as paper #1). This paper mainly found the effectiveness of antiparasitic ivermectin in reducing the risk of death in people infected with COVID-19 or high-risk groups through a meta-analysis of 15 randomized controlled trials. In terms of citation diffusion, paper #1 was cited more than 50 times by the academic community before it was “expressed as a concern”. For example, Boretti A (2022) suggests the best way to treat COVID-19 infections and indirectly treat Nigella sativa infections based on the results of paper #1 (cited as paper #2). Santin A D et al. (2021) cite paper #1 as one of the notable evidence in support of ivermectin's efficacy in reducing COVID-19 mortality (cited as paper #3). Paper #3 has also been cited 11 times to the present day, with an AAS of 17,121, having a significant social impact.

In terms of information diffusion, paper #1 has attracted widespread social attention since its publication, with 45,585 Twitter mentions and has been mentioned on blogs, mainstream media, and multiple types of communication platforms such as Wikipedia. Paper #1 was still being discussed by the public at the time of data collection for this study. Numerous related reports have led the public to believe in the effectiveness of ivermectin, and even to view it as a stopgap measure in the event of a vaccine shortage. However, as of now, health authorities such as the U.S. Food and Drug Administration (FDA) (2021) has recommended against using ivermectin for COVID-19 treatment outside of clinical trials, citing insufficient evidence of its efficacy and safety.

The veracity and reliability of Paper #1 have been questioned due to claims of data collection or reporting flaws in at least two of the data sources it incorporates. Specifically, one of the data samples that were the subject of the allegations was a paper published by Elgazzar A et al. (2020) based on the results of a clinical trial (notated as Paper #4) claiming that ivermectin reduced mortality from neocoronaryngitis by more than 90%. This paper was ultimately retracted by the preprint server Research Square due to possible plagiarism and data manipulation issues. After evaluating paper #1, the journal editors labeled the study "Expression of Concern" (Manu, 2022; Reardon, 2021) because they believed that the exclusion of questionable data, such as paper #4, might invalidate the study's results. Until the end of data collection in this study, the investigation of the allegations against the data sample of paper #1 remained inconclusive, which had a lasting negative effect on the meta-study and led to a cascade of negative impacts in subsequent academic research and social dissemination (shown in Figure 4).

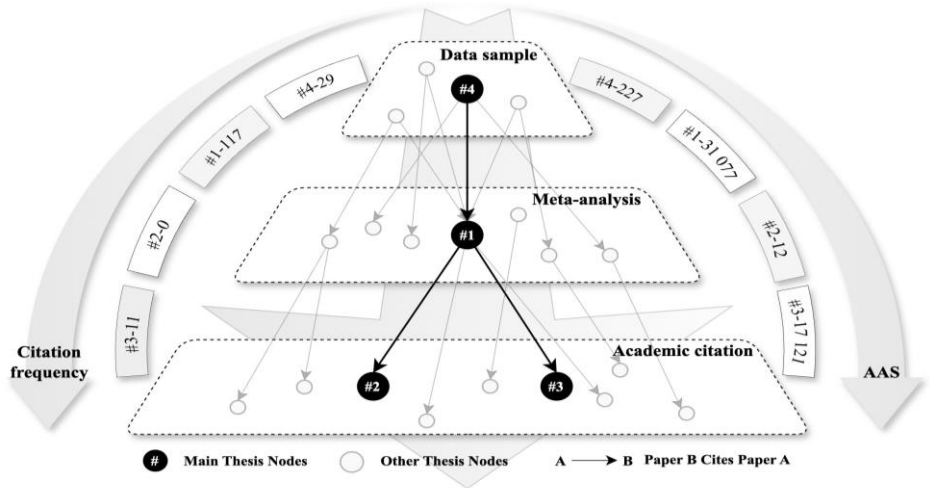


Figure 4. Illustration of the cascading impacts of COVID-19 retracted papers.

Discussion and conclusion

With the help of altmetric indicators, this study explores the social impact in terms of social attention, public dissemination, and policy making. Taking typical papers as an example, we analyze the cascade of impacts that may be triggered by COVID-19 retracted papers. The main findings and inspirations of this study are as follows.

The social impact of retracted papers is closely related to two landmark events: the publication and the retraction notice release. Among the hot papers that receive widespread public attention, the distribution of disciplines and countries shows significant concentration. Moreover, the release of retraction notices becomes a crucial window for researchers and the public to access detailed retraction information and respond to the potential negative effects. The delay in retraction, and the reasons for retraction play an important role in shaping its societal impact. However, a large number of papers with short retraction delays lack specific retraction reasons. In addition, in public discussions, not only the paper's research findings but also the retraction event and its reasons play a central role.

COVID-19 retracted papers exhibit a high average AAS, with a highly concentrated distribution of social impact, as 8% of the papers attract nearly 80% of the total attention. Notably, only a few problematic papers triggered retractions, but they caused widespread ripple effects, misleading subsequent research and public cognition. COVID-19 retracted papers attract differing attention from the academic community and the public. Through diverse media channels, they often reach broader audiences, with faster public engagement associated with longer-lasting discussions. Twitter shows the highest mention coverage, reflecting high overall dissemination, while intense news media coverage helps speed up the retraction process and improves corrective timeliness.

Moreover, COVID-19 retracted papers have a faster rate of impact on policy documents, averaging only 77.5 days from publication to citation. In policy documents, retracted papers — often withdrawn due to scientific errors — are primarily cited for practical purposes, with references typically made to their conclusions and discussions to support policy considerations. Fortunately, 87.2% of policy documents used standardized formats for paper mentions, which aids in automatic identification.

The characteristics of the paper, the delay in retraction and the reasons for retraction play an important role in the impact generated by the retracted paper

A large number of papers without a specific reason for retraction indicates that

journals should not only improve the timeliness of academic purification but also pay attention to the standardization of retraction notices and the normality of the retraction process. Furthermore, the gradual shift of social attention to the potential risk of retraction may continue for a long period after the retraction notice is released. In addition to the research results of the paper, the retraction event and the reasons for it also occupy an important position in the public discussion.

In purification, we should not only be highly alert to the risk of subsequent research due to scientific errors in papers but also resolutely prevent and crack down on academic misconduct. Especially in emergencies such as the COVID-19 epidemic, which urgently require rapid response and precise guidance from the scientific community, the maintenance of academic integrity is even more urgent. Retractions caused by academic misconduct may trigger a crisis of public trust in science and affect public perception of epidemic prevention measures (Yuan & Liu, 2024), which may have significant and difficult-to-eliminate negative social impacts.

The influence is highly concentrated, and there are differences between researchers and the public

Only a small number of papers that meet the urgent needs of the public can quickly gain a large amount of attention. Among these highly concerned papers, only very few may be problematic retracted papers. However, it is these papers that may cause great waves and trigger a sustained and extensive ripple effect. In turn, these papers misdirect the subsequent research direction and public cognition, leading to an overall information epidemic.

In addition, there are differences in the focus of academia and society on the COVID-19 retracted papers. The dialogue between researchers and the public on cutting-edge issues is not entirely equal, which affects the public's correct cognition of retracted papers. This suggests that researchers should take social responsibility in public health emergencies, pay attention to public needs and concerns, and give full play to the social value of scientific research by solving practical problems.

Considering the prompt and responsive social attention, news media and social media should cooperate to improve the timeliness of academic purification

Supported by open-access initiatives and social media platforms, the discussion of papers is no longer limited to scholars but has become the focus of a wider audience through various media. Khan H et al. (2022) have found that retracted papers may be more likely to receive extraordinary attention on social media platforms than non-retracted papers, especially for papers that the public can readily perceive as problematic. Similar to the findings of this study, COVID-19 non-retracted papers

exhibit “slow” dissemination characteristics (Mehra et al., 2020). This suggests that social media has a role to play in identifying “unreliable” papers, combating rumors, and popularizing science. Da Silva J et al. (2019) also mention that anonymous comments about academic misconduct are becoming commonplace on social media platforms such as Twitter. These comments tend to be quickly noticed and widely disseminated. In addition, the high attention of news media has a positive effect on accelerating the timeliness of retraction purification.

Therefore, the relevant regulatory authorities should fully allow the news media to guide mainstream public opinion, grasp the dissemination characteristics of social media, and prioritize the monitoring of academic achievements that are highly popular on these social media platforms. This can enhance the probability and speed of monitoring retraction through public opinion, thereby enhancing the timeliness of retraction purification, reaching the optimal effect of public memory correction during the period of social attention.

The dialogue between academia and policymakers has been strengthened, and papers mentioned in policy documents should be rigorously monitored

Compared with our study, Yu Houqiang et al. (2017) based on more than 90,000 policy document mentions collected from 2013 to 2016, found that less than 12% of papers were mentioned within 180 days, with an average delay of 4.5 years. This highlights that during COVID-19, multiple institutions, including medical, surveillance and scientific research, worked closely together and focused on relevant scientific research activities. This greatly strengthened the cooperation and dialogue between academics and policymakers and accelerated the speed of knowledge transformation in policy-making.

The policy documents primarily mention retracted papers for scientific error to utilize their conclusions and discussion sections, with minimal academic exploration and critique of the flawed papers. This may pose a potential threat to the scientific validity and efficacy of policy formulation. Furthermore, policy documents used standardized description formats when mentioning papers will improve the correct identification rate of automated processing of large batches of data and facilitate wider research. In addition, policymakers should rigorously monitor papers mentioned in policy documents, adequately assess the quality of papers, and make timely adjustments and updates in policy documents based on changes in the status of the papers. Therefore, it is important to maintain a constant focus on potentially defective papers and to mark their retracted status promptly. The negative effect of retracted papers must be minimized while taking full advantage of the authoritative information on academic achievements.

Limitations

This study has certain limitations, as it solely relies on altmetric indicators to analyze the impact of the papers, with content analysis focusing exclusively on the characteristics of mentions in policy documents. It does not comprehensively examine the mention characteristics of retracted papers across various dissemination platforms. Future research will adopt a media dissemination perspective to investigate the social impact of academic papers, with the goal of optimizing the dissemination model of research outputs, enhancing their visibility and recognition within the social domain, and developing a more robust and comprehensive system for evaluating the social impact of academic achievements.

References

Arroyo-Machado, W., Robinson-Garcia, N., & Torres-Salinas, D. (2022). A comprehensive dataset of the spanish research output and its associated social media and altmetric mentions (2016-2020). Data, 7(5), 59.

Bar-Ilan, J., & Halevi, G. (2018). Temporal characteristics of retracted articles. Scientometrics, 116(3), 1771–1783.

Boretti, A. (2022). Steroids induced black fungus infection in india during the May 2021 COVID-19 outbreak. Indian Journal of Otolaryngology and Head & Neck Surgery, 74(Suppl 2), 3216–3219.

Caulfield, T. (2020, April 27). Pseudoscience and COVID-19—We've had enough already. Nature. Retrieved May 27, 2024 from: https://www.nature.com/articles/d41586-020-01266-z.

Da Silva, J., & Dobránszki, J. (2019). A new dimension in publishing ethics: Social media-based ethics-related accusations. Journal of Information Communication & Ethics in Society, 17(3), 354–370.

Elgazzar, A., Hani, B., Youssef, S. A., Hafez, M., Moussa, H., & Eltaweel, A. (2020). WITHDRAWN: Efficacy and safety of ivermectin for treatment and prophylaxis of covid-19 pandemic. Research Square. DOI: 10.21203/rs.3.rs-100956/v3.

Geng, Y. (2020). New developments in media literacy in the context of information epidemics: Overseas experience and Chinese strategies [in Chinese]. News and Writing, 37(8), 13–23.

González-Betancor, S. M., & Dorta-González, P. (2023). Does society show differential attention to researchers based on gender and field? Journal of Informetrics, 17(4), 101452.

Guo, F., You, B., & Xue, J. (2016). Analysis on Transmission Characteristics and Influence of Altmetrics Hot Papers [in Chinese]. Library and Information Service, 60(15), 86–93.

Guo, L., & Zhou, Q. (2023). Academic Impact Evaluation of Disruptive Papers [in Chinese]. Digital Library Forum, 19(3), 19–27.

Huang, C., Huang, L., Wang, Y., Li, X., Ren, L., Gu, X., Kang, L., Guo, L., Liu, M., Zhou, X., Luo, J., Huang, Z., Tu, S., Zhao, Y., Chen, L., Xu, D., Li, Y., Li, C., Peng, L., … Cao, B. (2021). RETRACTED: 6-month consequences of COVID-19 in patients discharged from hospital: A cohort study. Lancet (London, England), 397(10270), 220–232.

Khan, H., Gupta, P., Zimba, O., & Gupta, L. (2022). Bibliometric and altmetric analysis of retracted articles on COVID-19. Journal of Korean Medical Science, 37(6), e44.

Li, J., Sun, L., Feng, S. rise in retractions in the life sciences literature during the pandemic years 2020 and 2021, He, P., & Zhang, Y. (2021). Social media communication of the scientific and technological literature in emergency under COVID-19. Library Hi Tech, 39(3), 796–813.

Liu, X., Sun, M., Xie, R., & Xiang, N. (2022). The Dissemination Power of COVID-19 Academic Achievements by Chinese Scholars Based on Altmetrics [in Chinese]. Documentation, Information & Knowledge, 39(3), 60–71.

Liu, X., Wang, C., Chen, D.-Z., & Huang, M.-H. (2022). Exploring perception of retraction based on mentioned status in post-retraction citations. Journal of Informetrics, 16(3), 101304.

Ma, L., Feng, L., Yuan, J., & Wang, L. (2023). Analysis and thoughts on the retracted papers published in SCI / SSCI journals sponsored by Chinese organizations [in Chinese]. Chinese Journal of Scientific and Technical Periodicals, 34(5), 584–592.

Manu, P. (2022). Expression of concern for Bryant a, Lawrie TA, Dowswell T, Fordham EJ, Mitchell S, Hill SR, Tham TC. Ivermectin for prevention and treatment of COVID-19 infection: A systematic review, meta-analysis, and trial sequential analysis to inform clinical guidelines. Am J Ther. 2021;28(4): E434-e460. American Journal of Therapeutics, 29(2), e232.

Mehra, M. R., Desai, S. S., Ruschitzka, F., & Patel, A. N. (2021). RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: A multinational registry analysis. The Lancet, 395(10240), 1820.

Reardon, S. (2021). Flawed ivermectin preprint highlights challenges of COVID drug studies. Nature, 596, 173–174.

Ren, C., & Yang, M. (2023). The Light at the End of the Tunnel: A Policiometric Analysis of Public Health Policies [in Chinese]. Library Tribune, 43(8), 31–42.

Ren, C., Yang, M., Li, K., Yang, G., & Lu, X. (2023). Analysis of Factors in Citing Scientific Papers in Policies against COVID-19 Pandemic [in Chinese]. Journal of the China Society for Scientific and Technical Information, 42(3), 341–353.

Resnik, D. B., Wager, E., & Kissling, G. E. (2015). Retraction policies of top scientific journals ranked by impact factor. Journal of the Medical Library Association, 103(3), 136–139.

Rubbo, P., Lievore, C., Biynkievycz Dos Santos, C., Picinin, C. T., Pilatti, L. A., & Pedroso, B. (2022). "Research exceptionalism" in the COVID-19 pandemic: An analysis of scientific retractions in Scopus. Ethics & Behavior, 33(5), 339–356.

Santin, A. D., Scheim, D. E., McCullough, P. A., Yagisawa, M., & Borody, T. J. (2021). Ivermectin: A multifaceted drug of Nobel prize-honoured distinction with indicated efficacy against a new global scourge, COVID-19. New Microbes and New Infections, 9(5), 100924.

Serghiou, S., Marton, R. M., & Ioannidis, J. P. A. (2021). Media and social media attention to retracted articles according to Altmetric. PLOS ONE, 16(5), e0248625.

Shen, D. (2022). Response of international publishing industry of medical journals in peer review and open science amid COVID-19 pandemic and its impact [in Chinese]. Chinese Journal of Scientific and Technical Periodicals, 33(8), 1046–1056.

Song, L., & Yang, W. (2023). Research on retracted papers involving ethical issues in science and technology [in Chinese]. Chinese Science Bulletin, 68(13), 1621–1625.

Sun, J., Wang, W., & Ding, Z. (2023). Academic characteristics of 253 retraction papers related to COVID-19 and implications [in Chinese]. Chinese Journal of Scientific and Technical Periodicals, 34(2), 241–248.

U.S. Food and Drug Administration. (2021, March 5). Ivermectin and COVID-19. Retrieved from https://www.fda.gov/consumers/consumer-updates/ivermectin-and-covid-19

Vainshelboim, B. (2021). RETRACTED: Facemasks in the COVID-19 era: A health hypothesis. Medical Hypotheses, 47(1), 110411.

Van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023—A new record. Nature, 624(7992), 479–481.

Xie, A., Yuan, L., & Wang, W. (2022). Reasons for the retraction of SCIE-indexed medical research articles by Chinese scholars [in Chinese]. Chinese Journal of Scientific and Technical Periodicals, 33(5), 554–560.

Yang, Z. (2020). Investigation and thinking on statements of retraction measures in Chinese academic journals [in Chinese]. Chinese Journal of Scientific and Technical Periodicals, 31(11), 1305–1310.

Yeo-Teh, N. S. L., & Tang, B. L. (2022). Sustained rise in retractions in the life sciences literature during the pandemic years 2020 and 2021. Publications, 10(3), 29.

Yin, Y., Gao, J., Jones, B. F., & Wang, D. (2021). Coevolution of policy and science during the pandemic. Science, 371(6525), 128–130.

Yu, H., Murat, B., Li, J., & Li, L. (2023). How can policy document mentions to scholarly papers be interpreted? An analysis of the underlying mentioning process. Scientometrics, 128(11), 6247–6266.

Yu, H., & Qiu, J. (2014). Theoretical research on stratifying and aggregating altmetric indicators [in Chinese]. Library Journal, 33(10), 13–19.

Yu, H., Xiao, T., Wang, Y., & Qiu, J. (2017). Study of Distribution Characteristics of Policy Documents Altmetrics [in Chinese]. Journal of Library Science in China, 43(5), 57–69.

Yuan, Z., & Jin, T. (2024). Characteristics of retracted papers published by top journals: A case study of Cell, Nature, and Science [in Chinese]. Chinese Journal of Scientific and Technical Periodicals, 35(02), 216–225.

Yuan, Z., & Liu, Y. (2024). Science and Technology Ethical Issues and Governance Paths in COVID-19 Retracted Papers [in Chinese]. Medicine & Philosophy, 45(3), 22–26.

https://doi.org/10.51408/issi2025_046

Structures of Authors' Collaboration at Young Universities

Nataliya Matveeva[1], Vladimir Batagelj[2]

[1]*nmatveeva@hse.ru*
HSE University, Moscow (Russia)

[2]*vladimir.batagelj@fmf.uni-lj.si*
Institute of Mathematics, Physics and Mechanics, Ljubljana (Slovenia)
University of Primorska, Andrej Marušič Institute, Koper (Slovenia)

Abstract

There are many studies devoted to university collaboration, but little is known about the existing structure of researchers' collaboration: which structures foster academic development and which do not. In our study, we analyze the co-authorship networks of eight leading young universities to investigate the collaboration structures of their researchers. We construct the corresponding co-authorship network for each university based on publication data from Scopus for the years 2017–2019. Our analysis includes two-mode university authorship networks, one-mode co-authorship networks, and subnetworks of authors who demonstrate the most productive collaboration. We found that the basic collaboration characteristics of leading young universities are quite similar. These universities exhibit a high level of collaboration, though the patterns of collaboration vary. The subnetwork of authors demonstrating the most productive collaboration reveals different structures based on the number of components and the geographic distribution of the authors. Our results highlight that collaboration is an important resource for leading young universities, but the collaboration structures of their authors differ significantly. Although overall collaboration is high, structural difference impact academic performance. Besides the prevalence of authors with certain types of affiliation, three collaboration models are identified: diverse collaboration, active intra-university collaboration, and active international collaboration. We discuss the risks associated with differing core compositions and propose policy recommendations based on our findings.

Introduction

Searching for the most effective methods and models of university development has been the subject of many studies. This question is especially relevant for young universities as they seek their path to academic success. Factors influencing university academic success have been studied at several levels: national (Heng et al., 2020), institutional environment (Altbach, 2009), and organizational (Amara et al., 2015; Goodall, 2009). Scientific collaboration is often cited as one of the main factors promoting academic excellence (Landry et al., 1996; Altbach & Salmi, 2011; Lim & Boey, 2014; Larivière et al., 2015). There are also recommendations for university policymakers to foster collaboration (Altbach, 2009; Abramo et al., 2009), and some young universities actively follow these recommendations (Costa, 2021).
In previous studies, various types of collaboration and their influence on academic performance have been examined. It has been shown that international collaboration positively affects the research performance of both individual scientists and universities (Ni & An, 2018; Matveeva et al., 2021) and that long-term collaboration has a greater impact on university development than short-term collaboration (Guskov et al., 2018; Altbach, 2009). Collaboration with industry has also been

found to positively influence university publication output (Bikard et al., 2019). Moreover, it is not only the type of collaborator that is important but also the position of scholars and universities within the academic network (Bordons et al., 2015; Chen et al., 2020). A central position in the network enhances access to information and other resources, facilitating their exchange. This, in turn, increases research activity and improves its quality. Another critical network characteristic is the probability of link formation (Ferligoj et al., 2015). The likelihood of link formation depends on many factors, including network structure, research policy, and institution-specific elements. Despite the extensive research on scientific collaboration, little is known about the structure of collaboration and its impact on universities' academic performance. Our work is devoted to the following question: Do leading young universities have the same collaboration strategies?

There are different approaches to analyzing the scientific collaboration of universities. Often, collaboration between universities is measured by the number of joint publications with various organizations and their distribution across different disciplines (Kotiranta et al., 2020; Matveeva et al., 2021). This approach emphasizes the intensity of collaboration and its preferences but does not reveal the impact of individual authors or the structure of collaboration. Another approach is fractional analysis (Batagelj, 2020; Demaine, 2022), which considers the impact of individual authors in collaboration. The fractional approach can be applied to both raw bibliometric data and network data. Co-authorship network analysis, on the other hand, explores the ways and channels for the transmission and dissemination of knowledge, thereby revealing the structure of collaboration (Mali et al., 2012; Matveeva & Ferligoj, 2020).

In our study, we apply co-authorship network analysis to investigate the collaboration structures of selected leading young universities. Our work addresses the following questions: Do leading young universities have the same collaboration structures? What parameters are similar, and which ones differ? To answer these questions, we analyze the co-authorship network of each university at two levels. First, we investigate 'full' non-normalized co-authorship networks to examine the general collaboration characteristics of the universities. We analyze both two-mode authorship networks and one-mode co-authorship networks. Then, from normalized co-authorship networks, we extract subsets of authors with the most productive collaboration (Ps-core). For this subset, we analyze the collaboration structure and the geographic distribution of authors.

The work is structured as follows: In the theoretical chapter, we describe the features of selected leading young universities and their collaboration strategies. The next chapter is devoted to the description of the data and methodology. In the chapter outlining the results, we provide an analysis of two types of co-authorship networks and the subset of authors with the most productive collaboration. In the final section, we discuss the findings and their limitations.

Features of leading young universities and factors influencing collaboration strategies

The term 'leading university' can have multiple interpretations, as leadership can be demonstrated in various areas such as research, teaching, the local labour market, the global academic market, and more. As usual, universities choose one or several niches and make efforts to take leading positions in them. To measure universities' activity and detect the leaders there are different World Universities Rankings. Each ranking has its methodology and procedure of inclusion, and there is a correlation between them (Robinson-Garcia et al., 2019). However, the influence of English-speaking countries is especially noticeable in the THE and QS Rankings (Moed & Moed, 2017), while ARWU is strongly biased towards US universities (Safón, 2013), and the Nature Index is biased towards natural science. The majority of rankings are biased towards research universities (Vernon et al., 2018).

For young universities, achieving a World Ranking is an important and desirable success indicator. Due to the lack of financial and reputation resources (Altbach & Salmi, 2011), young universities are compelled to search for additional sources to support their activities. Collaboration could be one such source. Indeed, several studies have identified collaboration with business and industry (Mok, 2013) and other leading universities (Lim & Boey, 2014) as factors that promote the academic success of young universities. However, little is known about the collaboration strategies of leading young universities, and we observe that these may vary (Crow, 2021). For some young universities their collaboration potential is not fully used (Khor & Yu, 2016).

In a work (Lancho-Barrantes & Cantu-Ortiz, 2021), for the sample of leading universities (not young) was found that top universities have strong research profiles, and some show more affinity among them than others. This result indicates that leading universities have different collaboration strategies and many factors may determine them. Due to the complex nature of scientific collaboration, choosing collaboration strategies is a relevant issue for both young and established universities. The complex nature of scientific collaboration is determined by the individual characteristics of scholars, institutional and organizational factors, country-specific elements, and a combination of all these aspects. Technology, intergovernmental programs, and policies are external factors influencing collaboration (Ribeiro et al., 2018; Larivière et al., 2015). On an individual level, motivation, research capability, and communication environment enhance collaboration (Zinilli et al., 2023). Academic culture, funding, institutional support, and the level of country institutionalization are institutional factors influencing scientific collaboration (Heng et al., 2020).

Country and culture-specific characteristics are also important in building a collaboration strategy. For instance, several studies have mentioned that global collaboration networks have a core-periphery structure, with Western countries at the core. Periphery countries face more challenges in research development, although their role is increasing year by year (Gazni et al., 2012; Gui et al., 2019). Often, the principle of cultural or geographic similarity is prevalent in countries' collaboration (Matveeva et al., 2022), meaning that collaboration more frequently

occurs between similar entities. Another crucial point is the influence of a country's research system on scientific collaboration. For example, for developing countries, building a strong collaboration system is an additional challenge due to the lack of specific institutes (Altbach, 2009; Heng et al., 2020).

In addition to external factors, the research profile of a university may also affect its collaboration. The positive effect of scientific collaboration on research productivity varies across different segments and institutional environments. The collaboration rate strongly correlates with the research discipline (Landry et al., 1996). On average, the number of authors per paper is higher in natural sciences and lower in social sciences and humanities (Larivière et al., 2006). In medicine and health sciences, research groups are most important, while international networks are most important in the natural sciences (Kyvik & Reymert, 2017).

Along with standing factors, many direct and indirect government research programs influence universities' scientific collaborations. Some government research programs have a strict focus on stimulating scientific collaboration. For example, a program in Japan encourages scientific collaboration to increase university visibility and attract international researchers and students (Ota, 2018). Yonezawa & Shimmi (2015) note the positive trend of Japanese universities' internationalization, although the impact is smaller than expected. In other programs, collaboration is not a priority but remains important. Universities may increase scientific collaboration to perform key indicators of government programs. The Russian Government Excellence Initiatives had a significant effect on universities' scientific collaboration (Matveeva & Ferligoj, 2020; Aldieri et al., 2020). Beyond government programs, universities have made efforts to internationalize themselves. For instance, Japanese universities increase the number of international programs for students and improve the English proficiency of their scientific staff (Ota, 2018). Singapore University NTU encourages partnerships with several leading overseas universities and multinational companies (Lim & Boey, 2014). Similarly, POSTECH University has actively developed a research network with top-class universities worldwide to become a world-class research institution (Altbach & Salmi, 2011).

There are various external and internal factors that affect the collaboration patterns of universities. With that, developing universities are in search of effective resources for growth, and collaboration can be this source. These universities often have a common mission and employ similar strategies to attain their goal: taking a leading position in the academic market. Analyzing the collaboration patterns of leading young universities can help us identify both common and distinctive collaboration characteristics. The common characteristics might represent effective practices that contribute to success, whereas unique characteristics could reflect national or institutional specificity of university.

Material and methods

In our study, we focus on young universities because they often have the same initial position. Collaboration may be their main resource due to the lack of other resources: human, financial, and reputation. As a measure of leadership, we use information about the position of the universities in different World University Rankings: Times

Higher Education (THE)[1], QS World University Rankings (QS)[2], Nature Index[3], Shanghai Ranking (ARWU)[4], and University Ranking by Academic Performance (URAP)[5]. We use several rankings to minimize bias toward concrete countries and disciplines (Robinson-Garcia et al. 2019). Nevertheless, these rankings are biased toward research universities, so our sample mostly represents young research universities.

We use the following steps for the sample formation:

1. Select the top 15% of the ranking for the analyzed years 2017-2019.
2. Choose universities that hold leading positions in at least three rankings.
3. Select young universities that were established after 1970.
4. Exclude merged universities as they are not really young.

After this procedure, we identified 8 leading universities, which also hold leading positions in the THE Young University Rankings. The analyzed universities are located in East and Southeast Asia, Europe, and Australia, and have either technical or general profiles (Table 1). For the observed period, NTU has the largest number of publications (25189), while UPF has the fewest (6054). The University of Sydney has the highest value of students per staff (43.40), and South Korean universities KAIST and POSTECH have the lowest (10.40 and 10.70 respectively). Moreover, UM has the highest share of international students (52%). This value can be explained by the university's location near the borders of several European countries. Korean universities demonstrate the lowest share of international students (4% in POSTECH and 9% in KAIST). The median place of the analyzed universities in the THE ranking is 125 (general ranking) and 9.5 (young universities ranking).

Table 1. The sample characteristics.

Name	Country	Year of establishment	Number of publications in Scopus 2017-2019	Number of students per staff*	Share of international students*	THE young & THE general rank in 2019	Dominant research fields since 2004**
Nanyang Technological University (NTU)	Singapore	1991	25189	16.70	0.28	3 & 47	Energy, Engineering, Computer Science
Hong Kong University of Science and Technology (HKUST)	Hong Kong SAR	1991	9225	23.60	0.31	1 & 56	Engineering, Materials Science, Computer Science
Korea Advanced Institute of Science and Technology (KAIST)	South Korea	1971	12543	10.40	0.09	6 & 96	Engineering, Materials Science, Chemical Engineering
Hong Kong Polytechnic University (PolyU)	Hong Kong SAR	1994	15155	27.80	0.25	15 & 129	Engineering, Business, Management and Accounting, Energy

[1] https://www.timeshighereducation.com/world-university-rankings
[2] http://www.topuniversities.com/
[3] https://www.nature.com/nature-index/institution-outputs/generate/all/global/all
[4] https://www.shanghairanking.com/
[5] https://urapcenter.org/

Pohang University of Science and Technology (POSTECH)	South Korea	1986	6179	10.70	0.04	8 & 152	Materials Science, Engineering, Energy
University of Technology Sydney (UTS)	Australia	1988	13094	43.40	0.36	13 & 160	Engineering, Computer Science, Environment Science
Maastricht University (UM)	Netherlands	1976	13338	15.80	0.52	11 & 121	Psychology, Neuroscience, Engineering
Pompeu Fabra University (UPF)	Spain	1990	6054	21.30	0.13	12 & 152	Social Science, Health Professions, Engineering

*According to the data of THE ranking for 2019
** According to Rankless company: https://www.rankless.org/about

For the analysis, we use all types of publications attributed to the universities' profiles in Scopus for the period 2017-2019. Before constructing the co-authorship networks, preliminary data preparation was conducted. We read and inspected the Scopus data, corrected parsing issues, removed duplicates, and extracted article IDs, author IDs, and authors' names. Based on the prepared publications dataset, we generated networks for each analyzed university. To analyze the scientific collaboration of the universities from various perspectives, we created three types of networks: the basic two-mode authorship network, the non-normalized one-mode co-authorship network, and the normalized subnetwork of the most productive authors. Since Newman's normalization does not account for loops in the co-authorship network (i.e., single-authored papers), we focus exclusively on authors who have collaboration with others. Thereby, our subnetwork includes only the authors with the most productive collaborations, where productivity defined by the number of papers.

Two-mode authorship network works-authors (WA)

For each university, we constructed two-mode authorship networks. These networks enable us to analyze the relationships between authors and their papers, linking the set of works with the set of authors. There is an arc (directed link) from work *p* to author *u* if and only if *u* is an author of work *p* (Batagelj & Cerinšek, 2013). We use the following network characteristics to analyze this network:

- number of rows and columns are the number of works and the number of authors respectively
- number of links counts the number of authorships
- maximum and average out-degree - the maximum and average number of authors per work
- maximum and average in-degree - the maximum and average number of works per author
- the distribution function of in - and out-degrees

Assume that the authorship network is described by the matrix WA. The projection $Co = WA^T*WA$ of the network WA produces a one-mode co-authorship network.

- Co[*a*, *b*] = number of works that authors *a* and *b* co-authored
- Co[*a*, *a*] = number of works co-authored by the author *a*

Co-authorship network (Co)

In this network, nodes represent authors, and links between them represent co-authored papers. The weight of a link corresponds to the number of co-authored papers. This network represents collaboration between scientists and is undirected. For this network, we calculate the following characteristics:

- number of nodes - the number of authors appearing in the university's bibliography co-authorship network
- number of links - the number of different pairs of co-authors
- number of components - the number of connected subgroups (CC)
- average degree - the average number of different co-authors that the author has
- distribution of (connected) components: size and proportion of the largest component LC = n(LC)/n, where n = number of nodes; link-proportion = m(LC)/m, m = number of links; proportion of isolated nodes IS = n(IS)/n
- size of the main core, n(MC) and the number of links in the main core, m(MC)

These characteristics allow us to analyze the size of connected and isolated groups and their share in the full network.

Subnetwork of authors with the most productive collaboration

In the standard co-authorship network, works with many co-authors are overrepresented. To make the publication output of authors comparable we normalized the analyzed networks. For the WA networks, we applied both Standard and Strict (Newman's) normalizations based on fractional approach (Batagelj, 2020). In the Standard normalized network n(WA), each row is divided by its degree and the author's self-collaboration (loops) is taken into account:

$$n(WA)[p, a] = WA[p, a]/\deg(p) \quad (1)$$

In Strict normalization loops are not consider:

$$n'(WA)[p, a] = WA[p, a]/(\deg(p)-1) \quad (2)$$

We multiplied two normalized WA networks to obtain a normalized network Ct' (the detailed procedure is described in (Maltseva & Batagelj, 2022)):

$$Ct' = D_0(n(WA)^T * n'(WA)) \quad (3)$$

where the function $D_0(M)$ sets the diagonal of a square matrix M to 0.

From the normalized Ct' network, we extracted a Ps-core. A Ps-core at level *t* is the maximal subset where each node's (author's) contribution (weighted degree = sum of author's link weights) in collaboration with authors within the core is greater or equal to the threshold *t* (Batagelj & Zaveršnik, 2011; Batagelj et al., 2014). For each analyzed university, we examined the distribution function of Ps-core values and chose the cut level that produced around 100 nodes (authors). The Ps-cores are nested – if $t_1 < t_2$ then $Ps(t_2) \subseteq Ps(t_1)$. Decreasing the level increases the size of the Ps-core, but no old Ps-core node/link is removed.

For the Ps-core subnetwork, we considered the following:

- number of nodes
- number of clusters (core's connected components)

- number of clusters of different size
- nodes property (geography of authors' affiliations)

Data cleaning and all computations were done using the programs R (R Core Team (2023)) and Pajek (http://mrvar.fdv.uni-lj.si/pajek/).

Results

Authorship network characteristics of the universities

We start the analysis from the two-mode authorship network Works – Authors (WA). This network represents the connection between authors and their works. We observe that NTU has the largest number of publications while UM has the highest number of authors (Table 2). HKUST shows outlier collaboration characteristics. On average, one author in this university has 32.47 works. This is a huge value. In contrast, at other universities, a scholar typically has 2-3 works. At HKUST, the maximum number of authors in one work is 5215 and the maximum number of works which one author has is 438. This suggests that HKUST actively participates in mega-science projects involving several thousands of authors. Beyond this outlier, we find that UPF and UM have the highest average number of authors per work (around 12 authors). With that, UPF demonstrates the lowest number of works that one author has on average (2.29 papers per author). Apart from HKUST, KAIST demonstrates high authors' productivity, with an average author having 3.53 papers. It should be noted that average values are sensitive to distribution function so this is more informative where the analyzed values have the same distribution function. Therefore, the distribution function of authors per work and works per author should be analyzed.

Table 2. Two-mode authorship networks of analyzed universities.

	POSTECH	KAIST	UM	NTU	UTS	HKUST	PolyU	UPF
Number of rows	6179	12543	13338	25192	13094	9225	15155	6054
Number of columns	15772	28006	66996	57779	32235	28108	32010	30715
Number of links	37487	98952	161754	152146	81505	912770	91058	70353
Average in-degree	2.38	3.53	2.41	2.63	2.53	32.47	2.84	2.29
Average out-degree	6.07	7.89	12.13	6.04	6.22	98.95	6.01	11.62
Max out-degree	382	1555	2582	1211	1211	5215	1211	1211
Max in-degree	169	105	122	188	192	438	248	117

We observe that the distributions of the number of works per number of authors of a work are not similar among the analyzed universities (Figure 1). Both HKUST and KAIST have a 'long tail' featuring works with several thousands of authors. UPF and

UM have a notable number of works containing between 100 and 1000 authors. POSTECH University, on the other hand, has very few multi-authored works. Despite these differences, the distributions of the number of works per number of authors across all universities indicate that the number of works decreases when the number of authors per paper exceeds five.

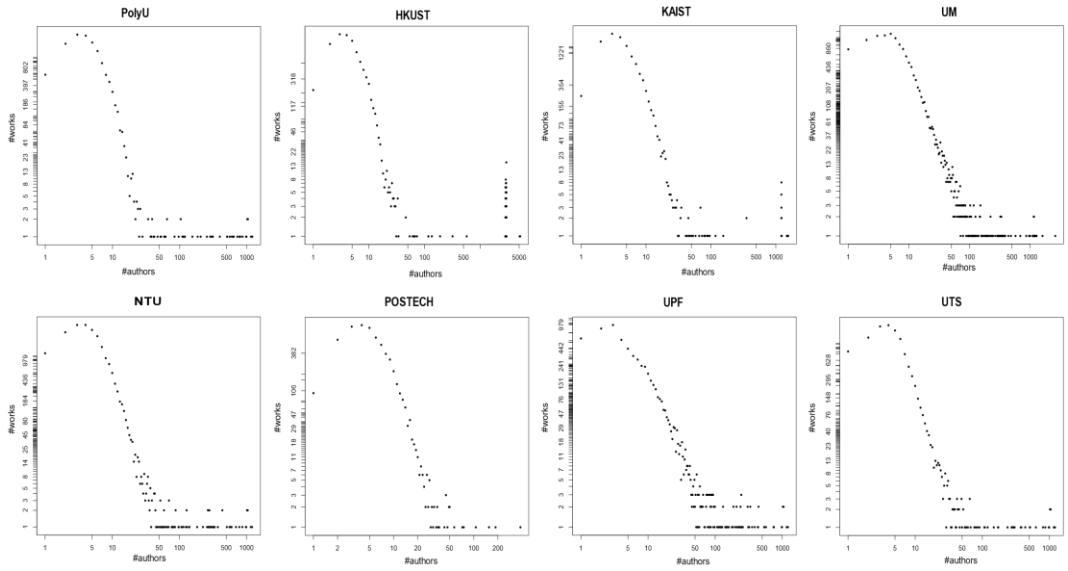


Figure 1. Distributions of the number of works per number of authors of a work for all universities.

When examining the distribution of the number of authors per number of works of an author, we observe comparable collaboration characteristics across seven universities, with HKUST being an outlier (Figure 2). The red lines indicate the 90% quantile of distribution. We find that in most of the analyzed universities, 90% of authors have between 4-7 works, except for HKUST, where this value is 149. There are also some outlier groups in UM, where many authors (more than 1,000) have published numerous papers. This can be attributed to the prevalence of multi-author works, typical in the High Energy and Particle Physics fields (Matveeva et al., 2021). Another important observation: analyzed universities have completely different numbers of authors with only one paper for 3 analyzed years. NTU does not have such authors. It seems that this university concentrates on research activity and authors have some obligations or stimulus to have more than 1 paper per 3 years. In UM, the opposite situation is observed: 43452 authors (from 66996 total) have only one work for 3 analyzed years. This university does not have a high number of students per staff (Table 1), so we may assume in UM there is no strong obligation to publish.

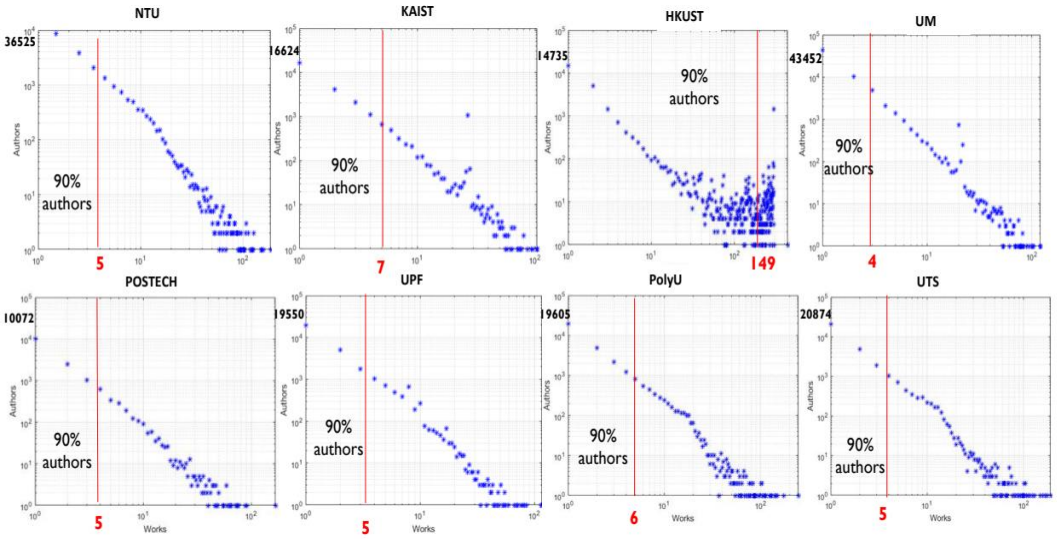


Figure 2. Distributions of the number of authors per number of works of an author for all universities.

We look at the co-authorship network of authors in universities' networks to analyze collaboration patterns between authors (Table 3). In this network, authors are nodes and links represent co-authorships - the weight of a link counts the number of publications that the linked authors co-authored. The analyzed universities are well-connected and have a comparable proportion of the largest component (the number of connected nodes divided by the number of nodes in the network). The proportion of links in the component is almost the same in all analyzed universities. Here we also observe some university-specific collaboration patterns. NTU has the highest number of connected authors and the number of connected components is two times higher than in other universities. With that, the number of direct co-authors (average degree centrality) at NTU is not high in comparison with the other universities (213.36 at NTU and 599.12 at UM). For this university, it is typical to collaborate in small separated groups. POSTECH has the lowest number of connected authors and degree centrality. On average, one author is connected with 55 co-authors in a network. This value is the lowest in the sample but still big. Such average value is explained by the presence of works with many authors (for example, 382 authors per work, see Table 2). For this university, it is typical that authors collaborate with one or several well-connected groups. Here we also observe unique collaboration patterns of HKUST: the huge value of degree centrality, although the number of connected groups in this university is not so high. These observations suggest that there are numerous local collaborations among authors and these collaborations occur within connected groups. The presence of a large connected component is typical for UM, while KAIST has the lowest number of isolated nodes — almost all nodes are connected.

Table 3. Co-authorship networks of analyzed universities.

	POSTECH	KAIST	UM	NTU	UTS	HKUST	PolyU	UPF
Number of nodes	15772	28006	66996	57779	32235	28108	32010	30715
Number of links	439464	4279814	20069338	6163927	4866335	45365272	4460480	6110717
Number of connected components (CC)	141	283	390	739	429	372	422	381
Average degree	55.72	305.63	599.12	213.36	301.92	3227.92	278.69	397.89
Number of isolated nodes (IS)	29	28	176	281	199	46	125	187
Proportion of the largest component (LC)	0.930	0.932	0.937	0.917	0.908	0.888	0.906	0.922
Link-proportion of the largest component (LC)	0.469	0.496	0.414	0.493	0.494	0.499	0.484	0.497
Share of isolated nodes (IS), %	0.18	0.10	0.26	0.49	0.61	0.16	0.39	0.61

Subnetworks of the most productive collaboration

In the previous section, we analyzed the structure of entire university networks, which consist of all authors mentioned in publications. However, often a significant portion of a university's publication output is produced by a limited number of researchers. Moreover, authors who are not affiliated with the university also contribute to the university's publication output. In this chapter, we analyze the collaboration structure of the authors who demonstrate the most productive collaboration. In the collaboration structure, we examine the number of clusters in the core, the level of authors' connection and the geography of their affiliations. Furthermore, it helps us understand the role of external authors in the university's publication output.

To observe the core, we extracted subsets of authors with the highest Ps-core values from Newman's normalized co-authorship networks. For each university, we decided to select a level of collaboration t that would produce the core of the size comparable across universities, approximately consisting of 100 authors. The authors within the core have the highest number of joined publications, calculated by taking into account the number of authors involved in each publication. We

extracted information about the affiliations of the authors in the core from their Scopus profiles. The main organization listed on the author's Scopus profile page was used to identify the authors' affiliation. In figures 3-10, world regions are represented by a color, while locations are indicated by symbols (Table 4).

Table 4. Notation of Word regions and Location code.

4a. World regions code

Yellow	East and Southeast Asia
Green	Europe
Red	North America
Blue	South Asia
Pink	Australia and Oceania
White	South America
Orange	Middle East
Purple	Central Asia

4b. Location code

At the selected university	Star
In the same country as the selected university	Rhombus
Other	Circle

In the core of PolyU, we observe one large cluster and several clusters of average size. Moreover, the majority of authors are from the same region: East and Southeast Asia (Figure 3). Many authors are from the same university. In this core, domestic collaboration within the country prevails, and international collaboration with other regions is minimal. Often, clusters (core components) include one or several authors from PolyU who collaborate with scholars from other organizations.

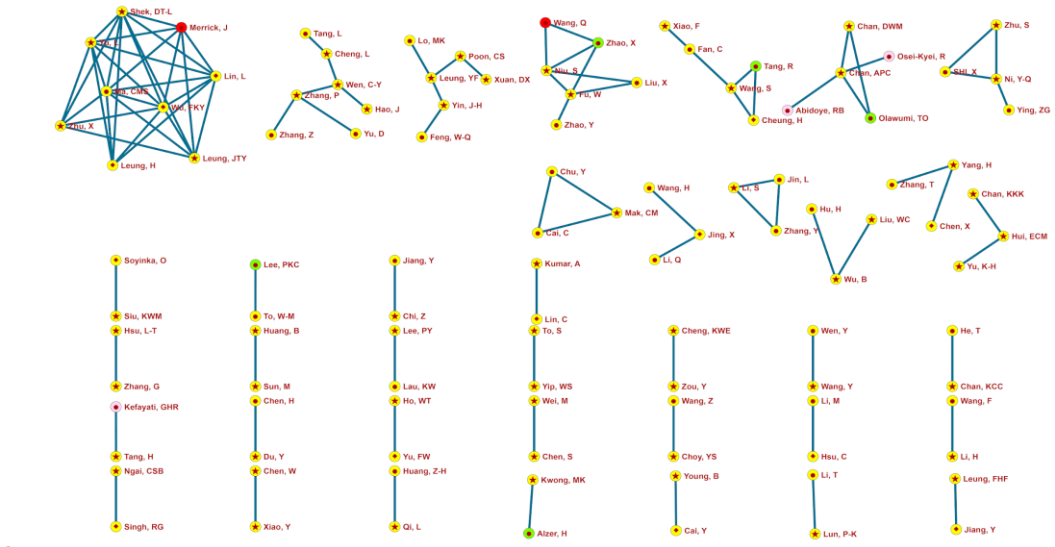


Figure 3. Ps-core of PolyU at level 4.03 (110 nodes).

South Korea's POSTECH University has a core structure similar to PolyU: there is one large cluster and several clusters of average size. There are a few small clusters with only two authors. In the POSTECH core, there are also numerous collaborations within the university and the country (Figure 4). Large clusters primarily consist of scholars who work in POSTECH or other South Korean organizations. There are a few collaborations with regions abroad in small clusters, including South Asia, North America, and Europe. However, collaboration with scholars from other East and Southeast Asia countries (denoted by yellow nodes with circles) is not typical for the POSTECH core.

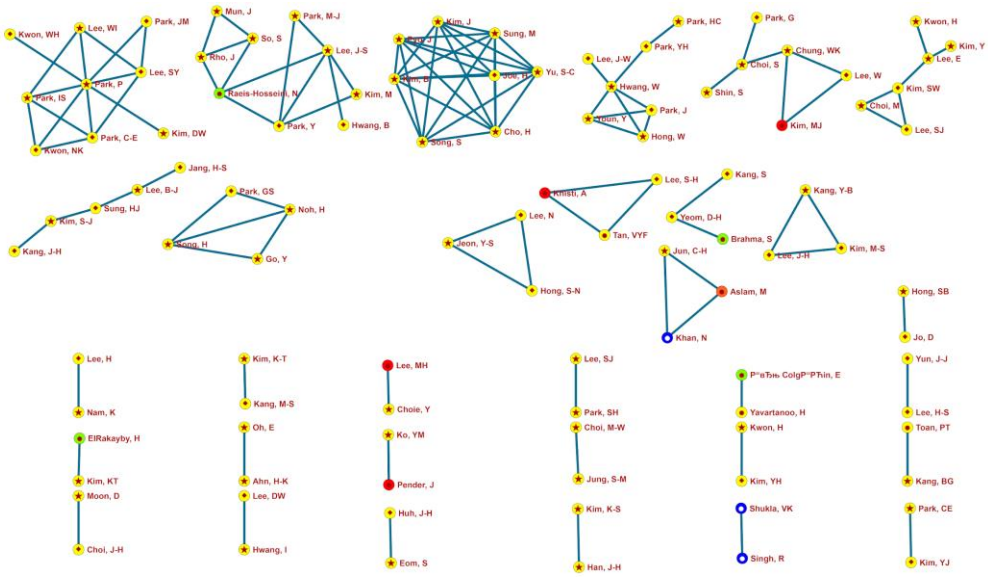


Figure 4. Ps-core of POSTECH at level 2.2 (107).

In contrast, HKUST University (Figure 5) has a slightly different core structure compared to the previous two universities: there are several clusters of average size and many small clusters. We observe collaborations within the university and region in average and large groups, where authors from HKUST collaborate with others. In small clusters with 2-3 authors, collaborations with scholars from North America are often seen. HKUST's collaboration within the country is very weak.

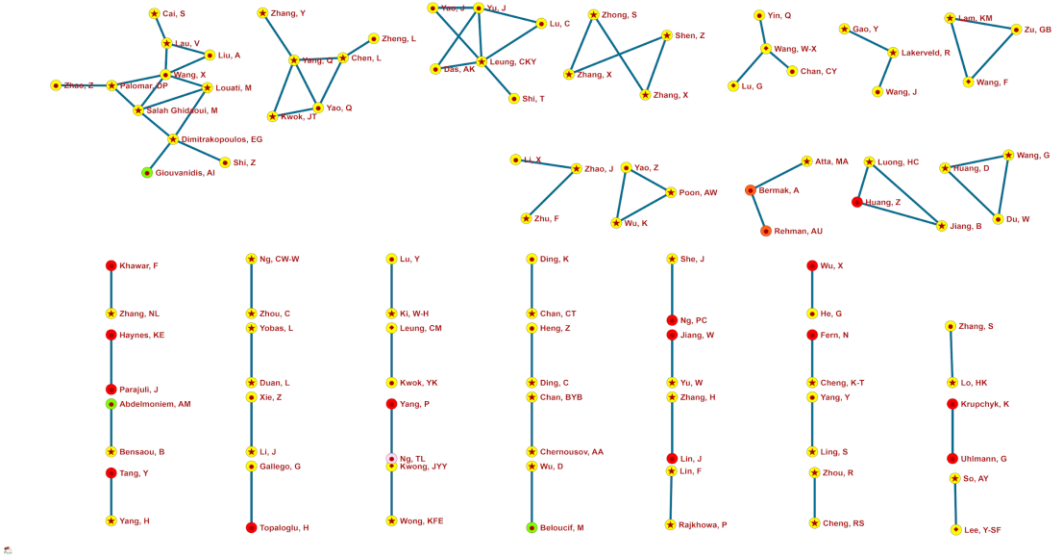


Figure 5. Ps-core of HKUST at level 1.36 (106 nodes).

The structure of the KAIST core is similar to HKUST as well as UM, UPF, and UTS. Here also there are several clusters of average size and many small clusters (Figure 6). With that, in KAIST there is intense domestic collaboration (inside the university and country). Often, the clusters consist of several authors from KAIST University. International collaboration is minimal.

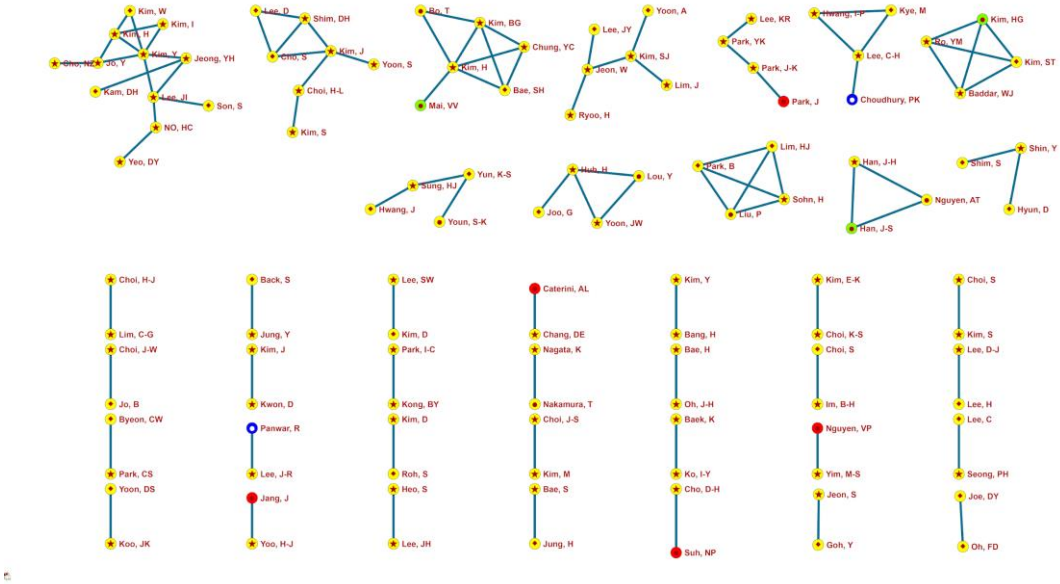


Figure 6. Ps-core of KAIST at level 3.34 (117).

In the UTS core, there are collaborations within the university, within the country, and abroad with different regions (Figure 7). Notably, collaboration with countries from the same region is absent in the UTS core. There is one cluster consisting solely of scholars from East and Southeast Asia. Scholars in this cluster have strong ties with each other and minimal connections (less than the chosen level of 3.33) with scholars from UTS.

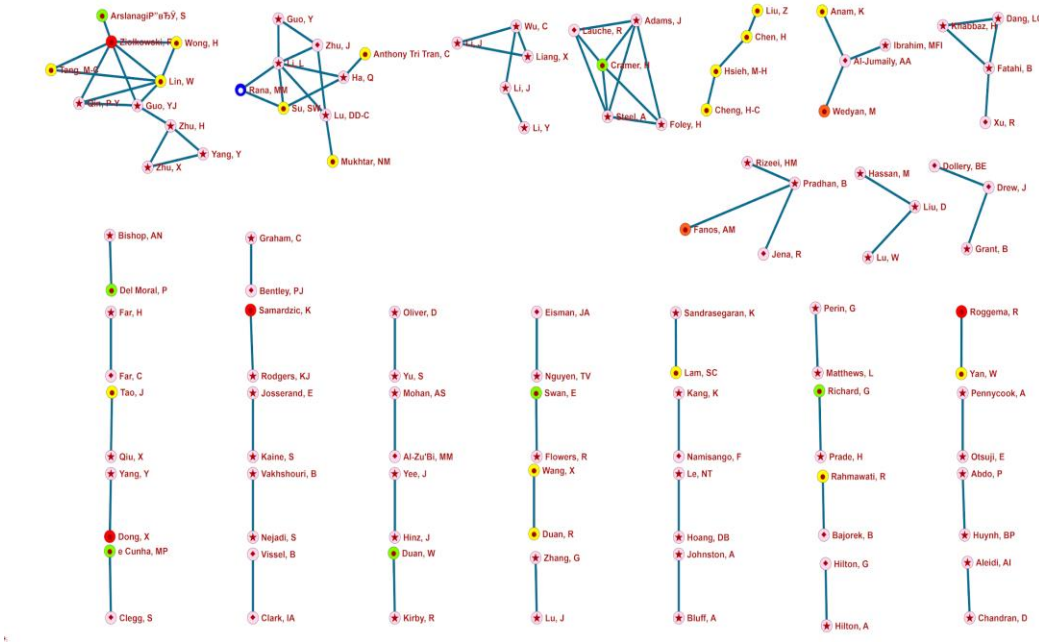


Figure 7. Ps-core of UTS at level 3.13 (111 nodes).

In the UM core, most collaborations are from the same region (Europe), both within and outside the country (Figure 8). Scholars from UM rarely collaborate with each other. Often, clusters include one or several individuals from UM who collaborate with scholars from other organizations. The largest clusters consist of one author from UM and many authors from abroad. Collaboration within the country and outside the region is poor.

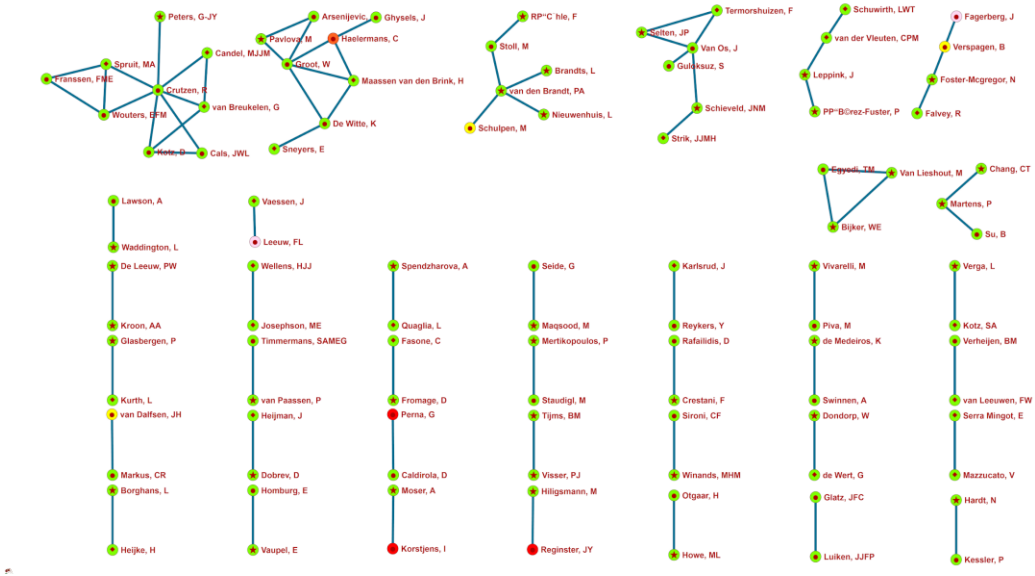


Figure 8. Ps-core of UM at level 2.33 (103 nodes).

In the UPF core, there are numerous collaborations within the region (Figure 9). However, compared to UM (also a European university), UPF has more collaborations within the university and country. UPF also collaborates with scholars from different regions within Europe. The diversity of foreign partners is higher than in UM, with authors from various regions present.

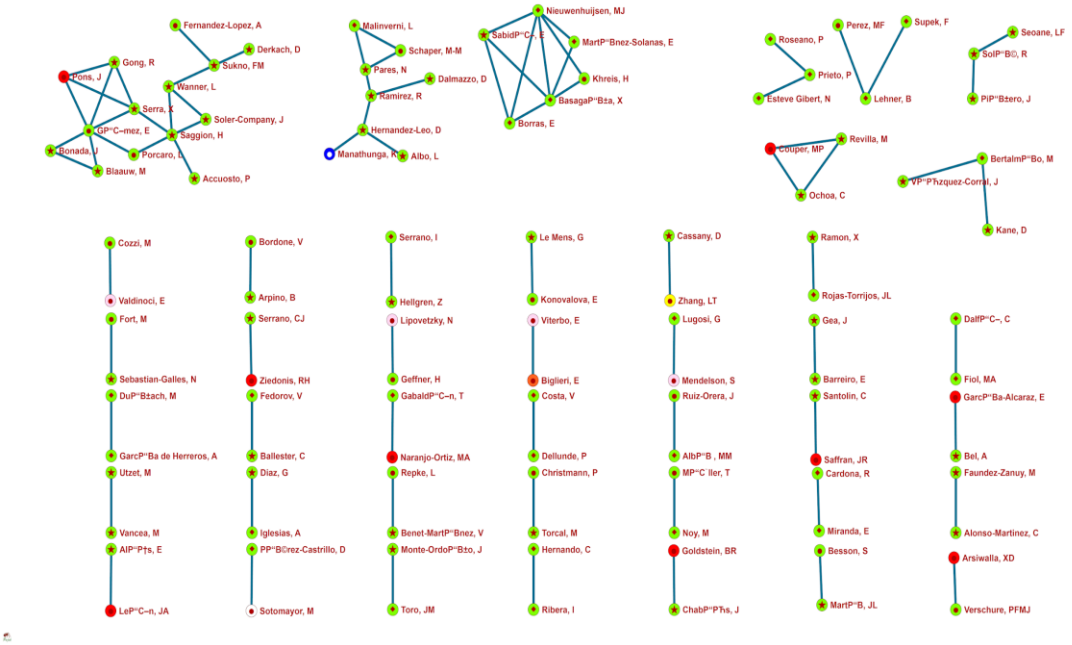


Figure 9. Ps-core of UPF at level 2.02 (111 nodes).

The structure of the core at NTU differs from that of the other universities analyzed. There are numerous small-sized clusters, while a large cluster is noticeably absent (Figure 10). In NTU, collaboration from different regions is present. The core consists of clusters from small and average groups, where one author is from NTU and another from other organizations. There are many collaborations out of the region, particularly with Europe and North America. Collaboration between scholars from NTU is almost absent, the core of the most productive authors working in collaboration is mostly located outside the university.

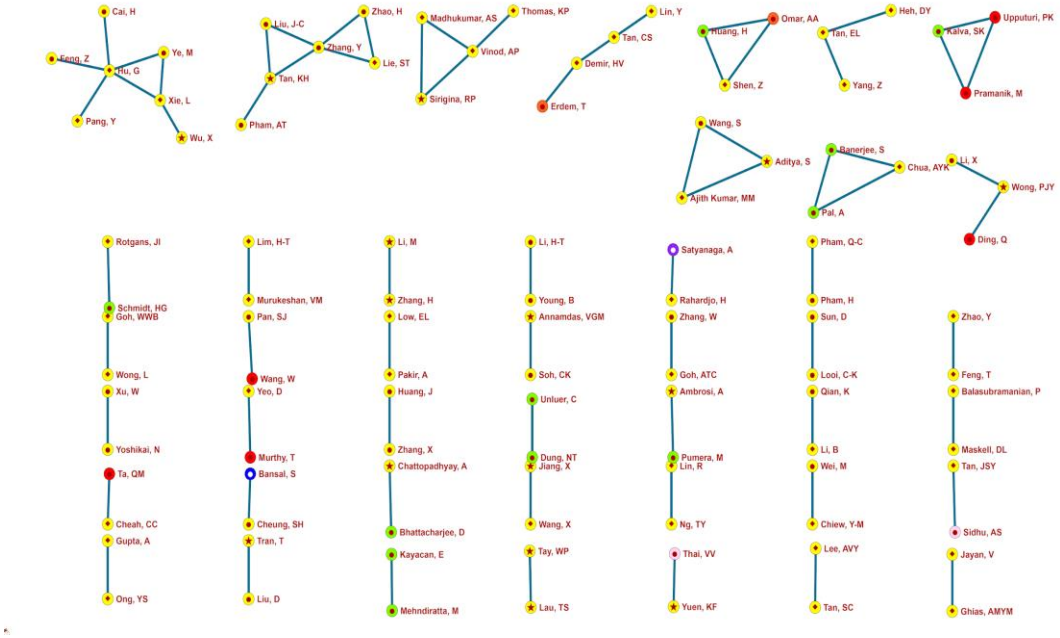


Figure 10. Ps-core of NTU at level 4.33 (107 nodes).

Among the analyzed universities, NTU demonstrates the most distinct characteristics of the core, as shown in Table 5. Core of this university has the highest level of authors' connection and the highest number of separated groups. This suggests that the authors at this university work in many separate groups, producing a substantial number of papers. While this structure is relatively stable and doesn't rely on specific authors, it poses certain risks given that only 15.60% of authors in the core are affiliated with NTU. HKUST also demonstrates a high share of foreign authors in the core, but these authors are not as connected as in NTU. POSTECH presents another core structure: it has a low number of separate groups, and the proportion of foreign authors is small (14.02%). These foreign authors are primarily from the regions of Europe and North America. KAIST demonstrates similar characteristics as POSTECH. The remaining universities in the sample have a relatively similar core structure, represented by the number of separate groups and the proportion of foreign authors.

Table 5. Characteristics of collaboration within the universities' core of the most productive authors working in collaboration.

	Country of the university	Core level t (productivity)	Number of clusters	Share of authors in the core affiliated with the university, %	Share of foreign authors in the core, %	Dominant region of foreign authors in the core
POSTECH	South Korea	2.20	32	52.34	14.02	Europe; North America
KAIST	South Korea	3.34	40	59.83	13.68	East and Southeast Asia; North America
UM	Netherlands	2.33	38	32.04	38.83	Europe
NTU	Singapore	4.33	44	15.60	44.04	East and Southeast Asia; Europe
UTS	Australia	3.13	40	59.46	27.93	East and Southeast Asia
HKUST	Hong Kong SAR	1.36	39	53.77	44.34	North America; East and Southeast Asia
PolyU	Hong Kong SAR	4.03	38	53.64	36.36	East and Southeast Asia
UPF	Spain	2.02	42	45.05	30.63	Europe; North America

We can conclude that, based on the number and size of core clusters, the analyzed universities can be categorized into two groups:

A) One large group and a few small groups: POSTECH, PolyU.

B) Several groups of average and small sizes: KAIST, UTS, UPF, POSTECH, HKUST, NTU

The authors from the core are affiliated with different organizations. Sometimes, these authors do not work at the universities being analyzed. Moreover, they do not connect with the authors from analyzed universities at selected collaboration level *t*. The weights of some links can be smaller than *t* because the property considered is the weighted degree - the sum of links with a given end-node. According to the prevalence of authors with certain types of affiliation, we can further distinguish the analyzed universities into three groups:

1. Diverse collaboration: This includes PolyU, HKUST, UPF, and UM. Both the proportion of foreign authors and authors affiliated with the university exceed 30%.
2. Active intra-university collaboration: This group includes KAIST, UTS, and POSTECH. Conversely, the proportion of authors affiliated with the university is 30% or more, while the proportion of foreign authors is less than 30%.
3. Active international collaboration: NTU is in this category. At the core, the proportion of foreign authors is 30% or more, and the proportion of authors who have worked at the university is less than 30%.

Discussion and conclusion

Scientific collaboration is often perceived as an additional resource for development, as it enhances research skills and provides access to new knowledge and equipment. Several leading universities have highlighted collaboration as a key component of their development strategies (Altbach & Salmi, 2011; Lim & Boey, 2014). In this study, we examine the common collaboration characteristics of eight leading young universities, which may contribute to their academic success. Co-authorship network analysis was used to investigate both the general collaboration characteristics of these universities and the relationships between individual authors. We examined three types of university networks to analyze the relationships between authors and their papers, between authors themselves, and among the most productive authors. Both the absolute value of network characteristics and the distribution function of certain characteristics were analyzed. Our results reveal that there are some common features but also many university-specific characteristics of collaboration.

In absolute terms, the publication output of the analyzed universities varies significantly, but the average productivity of authors is identical across almost all universities. Despite this, the universities have different numbers of authors who published only one paper during the three years analyzed. For instance, NTU has no authors who published just one paper, while at UM and UTS, such authors constitute 65% of the total. We observed that all the analyzed universities demonstrate a high level of scientific collaboration, with an average of between 6 and 98 authors per paper. The majority of the analyzed universities (7 from 8) have comparable general collaboration characteristics: average works per author, average authors per work, maximum authors of a work, and maximum works of an author. One university, HKUST, significantly diverges in these characteristics, demonstrating atypical values (for example, an average of 98.95 authors per paper). This university actively participates in multi-author collaboration works that involve several thousand authors, making the average value sensitive to such outliers.

All the universities' networks are well-connected and maintain a comparable proportion of largest component, calculated as the number of nodes in the largest component divided by the total number of nodes in the network. With that, analysis of collaboration structure reveals many university-specific collaboration patterns. For example, for NTU it is typical to collaborate in small separated groups, and this is a tight collaboration with many links. In UM, the opposite situation is observed: many authors in the network are connected but the connectivity is weak. KAIST demonstrates high connectivity among the authors in the network, with a minimal proportion of isolated nodes. HKUST has high local centrality of some parts of the network provided by multi-author's work. Specific collaboration structure can be explained by the dominance of certain research fields with established collaboration patterns. According to our preliminary analysis of the universities' profiles (Table 1), we observe that the mentioned universities have similar research profiles. Therefore, the observed differences can be related with the organization-specific or institution-specific characteristics of the universities.

In the final stage of our analysis, we examine the structure of the authors with the most productive collaborations and the geography of their affiliations. We observe that the collaboration structure of the authors in the analyzed universities differs a lot. Based on the number

and on the size of clusters in the core, the analyzed universities can be divided into two groups: one big cluster and a few numbers of small clusters (POSTECH, PolyU), several clusters of average and small sizes (KAIST, UTS, UPF, POSTECH, HKUST, NTU). We observed that universities vary in their core composition. Here, we distinguish three groups: those who have diverse collaboration (PolyU, HKUST, UPF, and UM), those oriented towards international collaboration (NTU), and those oriented towards intra-university collaboration (KAIST, UTS, and POSTECH).

Therefore, we observe that leading young universities actively use collaboration as a source for their development. On average, authors in these universities actively collaborate with other researchers from different regions. Furthermore, our study reveals marked differences in the collaboration structures of universities, depicting various collaboration patterns. Our study has several limitations. Firstly, we do not account for the thematic profiles of the analyzed publications, as network attributes and collaboration patterns can differ across various scientific disciplines. Secondly, the sensitivity of the collaboration structure to the chosen Ps-core level is another limitation; the structure of the core can vary based on this selection. The level of Ps-core characterizes the productivity of authors, so its selection depends on the research focus. Thirdly, we use the Scopus database, which does not fully cover local journals, especially those published in national languages (Vera-Baceta et al., 2019). Further research could take these limitations into account and focus on analyzing universities' decisions regarding collaboration strategies, specifically identifying the factors that influence these decisions.

We conclude that there is no single typical collaboration model for a leading young university; each university employs its own collaboration strategy, motivated by its vision, resources, and abilities. For instance, Altbach & Salmi (2011) provide some explanation about HKUST's strategies: ‘HKUST’s most important success factor was the recruitment of outstandingly talented scholars and scientists. The university recruits this caliber of academic staff from among the senior scholar generation of the Chinese diaspora. HKUST recruited heavily from this vast pool of talented academics born in Taiwan, China, or mainland China and trained overseas mostly at U.S. universities, something that the other universities in Hong Kong were less inclined to do at that time.’ Another finding is about POSTECH: ‘This university envisaged itself as a university offering excellence in education and research to Korean students who, thus, would not need to study abroad. To reach its goal, POSTECH developed a research network with top-class universities worldwide.’.

Our findings have some policy implications. Scientific collaboration, especially with leading research centres, is vital for knowledge exchange and university development. However, the structure of such collaboration can be unbalanced and may present risks for the university. For example, a high proportion of foreign authors in the core indicates that a significant portion of the publication output is produced outside the university. This model is sensitive to institutional or country-specific risks, where barriers to collaboration may arise, and is not inherently stable. Another example is that a collaboration network with a single big cluster of connected authors is less stable than a network with separate, diverse groups. A single cluster can be disrupted for various reasons, potentially triggering a domino effect within it. In contrast, a structure with multiple groups tends to be more stable. Policymakers should consider not only the quality and intensity of collaboration but

also the degree of staff involvement and the balance between foreign and intra-university collaboration. Achieving this balance can be facilitated by implementing targeted programs for various research groups. The first step in this direction is to assess the proportions of foreign and domestic authors and their collaboration structures. Our work is an exploratory analysis aimed at gaining insight into the factors influencing collaboration. We plan to continue the analysis based on a larger number of high-rank universities, both old and new.

Acknowledgments

All computations were performed using the program for large network analysis and visualization Pajek and the statistical programming system R. This work is supported in part by the Slovenian Research Agency (research program P1-0294 and research projects J5-2557 and J5-4596), research program CogniCom (0013103) at the University of Primorska, and prepared within the framework of the COST action CA21163 (HiTEc) and the HSE University Basic Research Program.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2009). Research collaboration and productivity: is there correlation? Higher Education, 57, 155-171. https://doi.org/10.1007/s10734-008-9139-z

Altbach, P. G. (2009). Peripheries and centers: Research universities in developing countries. Asia Pacific Education Review, 10, 15-27. https://doi.org/10.1007/s12564-009-9000-9

Altbach, P. G., & Salmi, J. (Eds.). (2011). The road to academic excellence: The making of world-class research universities. The World Bank. https://doi.org/10.1596/978-0-8213-8805-1

Aldieri, L., Kotsemir, M. N., & Vinci, C. P. (2020). The effects of collaboration on research performance of universities: An analysis by federal district and scientific fields in Russia. Journal of the Knowledge Economy, 11, 766-787. https://doi.org/10.1007/s13132-018-0570-9

Amara, N., Landry, R., & Halilem, N. (2015). What can university administrators do to increase the publication and citation scores of their faculty members? Scientometrics, 103, 489-530. https://doi.org/10.1007/s11192-015-1537-2

Batagelj, V. (2020). On fractional approach to analysis of linked networks. Scientometrics, 123(2), 621-633. https://doi.org/10.1007/s11192-020-03383-y

Batagelj, V., Doreian, P., Ferligoj, A., & Kejžar, N. (2014). Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution (Vol. 2). John Wiley & Sons. https://doi.org/10.1002/9781118915370

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. Scientometrics, 96(3), 845-864. https://doi.org/10.1007/s11192-012-0940-1

Batagelj, V., & Zaveršnik, M. (2011). Fast algorithms for determining (generalized) core groups in social networks. Advances in Data Analysis and Classification, 5(2), 129-145. https://doi.org/10.1007/s11634-010-0079-y

Bikard, M., Vakili, K., & Teodoridis, F. (2019). When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity. Organization Science, 30(2), 426-445. https://doi.org/10.2139/ssrn.2883365

Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. Journal of Informetrics, 9(1), 135-144. https://doi.org/10.1016/j.joi.2014.12.001

Chen, K., Zhang, Y., Zhu, G., & Mu, R. (2020). Do research institutes benefit from their network positions in research collaboration networks with industries or/and universities? Technovation, 94, 102002. https://doi.org/10.1016/j.technovation.2017.10.005

Crow, J. M. (2021). Young universities forge new paths to success. Nature, 600(7888), S6-S7. https://doi.org/10.1038/d41586-021-03630-z

Demaine, J. (2022). Fractionalization of research impact reveals global trends in university collaboration. Scientometrics, 127(5), 2235-2247. https://doi.org/10.1007/s11192-021-04246-w

Ferligoj, A., Kronegger, L., Mali, F., Snijders, T. A., & Doreian, P. (2015). Scientific collaboration dynamics in a national scientific system. Scientometrics, 104, 985-1012. https://doi.org/10.1007/s11192-015-1585-7

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. Journal of the American Society for Information Science and Technology, 63(2), 323-335. https://doi.org/10.1002/asi.21688

Goodall, A. H. (2009). Highly cited leaders and the performance of research universities. Research Policy, 38(7), 1079-1092. https://doi.org/10.1016/j.respol.2009.04.002

Gui, Q., Liu, C., & Du, D. (2019). Globalization of science and international scientific collaboration: A network perspective. Geoforum, 105, 1-12. https://doi.org/10.1016/j.geoforum.2019.06.017

Guskov, A. E., Kosyakov, D. V., & Selivanova, I. V. (2018). Boosting research productivity in top Russian universities: the circumstances of breakthrough. Scientometrics, 117(2), 1053-1080. https://doi.org/10.1007/s11192-018-2890-8

Heng, K., Hamid, M., & Khan, A. (2020). Factors influencing academics' research engagement and productivity: A developing countries perspective. Issues in Educational Research, 30(3), 965-987. https://doi.org/10.3316/informit.465283943914964

Khor, K. A., & Yu, L. G. (2016). Influence of international co-authorship on the research citation impact of young universities. Scientometrics, 107, 1095-1110. https://doi.org/10.1007/s11192-016-1905-6

Kotiranta, A., Tahvanainen, A., Kovalainen, A., & Poutanen, S. (2020). Forms and varieties of research and industry collaboration across disciplines. Heliyon, 6(3). https://doi.org/10.1016/j.heliyon.2020.e03404

Kyvik, S., & Reymert, I. (2017). Research collaboration in groups and networks: differences across academic fields. Scientometrics, 113, 951-967. https://doi.org/10.1007/s11192-017-2497-5

Landry, R., Traore, N., & Godin, B. (1996). An econometric analysis of the effect of collaboration on academic research productivity. Higher education, 32(3), 283-301. https://doi.org/10.1007/BF00138868

Lancho-Barrantes, B. S., & Cantu-Ortiz, F. J. (2021). Quantifying the publication preferences of leading research universities. Scientometrics, 126(3), 2269-2310. https://doi.org/10.1007/s11192-020-03790-1

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology, 66(7), 1323-1332. https://doi.org/10.1002/asi.23266

Lim, C. H., & Boey, F. (2014). Strategies for academic and research excellence for a young university: perspectives from Singapore. Ethics in Science and Environmental Politics, 13(2), 113-123. https://doi.org/10.3354/esep00139

Mali, F., Kronegger, L., Doreian, P., & Ferligoj, A. (2012). Dynamic scientific co-authorship networks. Models of science dynamics: Encounters between complexity theory and information sciences, 195-232. https://doi.org/10.1007/978-3-642-23068-4_6

Maltseva, D., & Batagelj, V. (2022). Collaboration between authors in the field of social network analysis. Scientometrics, 127(6), 3437-3470. https://doi.org/10.1007/s11192-022-04364-z

Matveeva, N., & Ferligoj, A. (2020). Scientific collaboration in Russian universities before and after the excellence initiative Project 5-100. Scientometrics, 124(3), 2383-2407. https://doi.org/10.1007/s11192-020-03602-6

Matveeva, N., Sterligov, I., & Yudkevich, M. (2021). The effect of Russian University Excellence Initiative on publications and collaboration patterns. Journal of Informetrics, 15(1), 101110. https://doi.org/10.1016/j.joi.2020.101110

Moed, H. F., & Moed, H. F. (2017). A comparative study of five world university rankings. Applied Evaluative Informetrics, 261-285. https://doi.org/10.1007/978-3-319-60522-7_18

Mok, K. H. (2013). The quest for an entrepreneurial university in East Asia: impact on academics and administrators in higher education. Asia Pacific Education Review, 14, 11-22. https://doi.org/10.1007/s12564-013-9249-x

Ni P., An X. Relationship between international collaboration papers and their citations from an economic perspective //Scientometrics. – 2018. – Т. 116. – №. 2. – С. 863-877. https://doi.org/10.1007/s11192-018-2784-9

Ota, H. (2018). Internationalization of higher education: Global trends and Japan's challenges. Educational Studies in Japan, 12, 91-105. https://doi.org/10.7571/esjkyoiku.12.91

Ribeiro, L. C., Rapini, M. S., Silva, L. A., & Albuquerque, E. M. (2018). Growth patterns of the network of international collaboration in science. Scientometrics, 114(1), 159-179. https://doi.org/10.1007/s11192-017-2573-x

Robinson-Garcia, N., Torres-Salinas, D., Herrera-Viedma, E., & Docampo, D. (2019). Mining university rankings: Publication output and citation impact as their basis. Research Evaluation, 28(3), 232-240. https://doi.org/10.1093/reseval/rvz014

Safón, V. (2013). What do global university rankings really measure? The search for the X factor and the X entity. Scientometrics, 97, 223-244. https://doi.org/10.1007/s11192-013-0986-8

Vera-Baceta, M. A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. Scientometrics, 121(3), 1803-1813. https://doi.org/10.1007/s11192-019-03264-z

Vernon, M. M., Balas, E. A., & Momani, S. (2018). Are university rankings useful to improve research? A systematic review. PloS One, 13(3), e0193762. https://doi.org/10.1371/journal.pone.0193762

Yonezawa, A., & Shimmi, Y. (2015). Internationalization: challenges for top universities and government policies in Japan. Higher Education, 70(2), 173–186. https://doi.org/10.1007/s10734-015-9863-0

Zinilli, A., Pierucci, E., & Reale, E. (2023). Organizational factors affecting higher education collaboration networks: evidence from Europe. Higher Education, 1-42. https://doi.org/10.1007/s10734-023-01109-6

https://doi.org/10.51408/issi2025_042

Study on the Differences Between Journal Papers and Conference Papers in the Frontier of Basic Research: Taking the Terahertz Field as an Example

Liu Hao[1], Chen Yunwei[2], Zhang Biao[3]

[1]*liuh@clas.ac.cn*, [2]*chenyw@clas.ac.cn*, [3]*zhangbiao231@mails.ucas.ac.cn*
National Science Library(Chengdu), Chinese Academy of Sciences, Chengdu, Sichuan (China)

Abstract

Analyzing the discrepancies in the content of journal papers and conference papers in the frontier of basic research is beneficial for a comprehensive understanding of the characteristics and patterns of basic research development. Retrieve data from the Web of Science (WoS) database in the frontier of basic research. Employ a comprehensive approach using bibliometrics, social network analysis, and text mining methods to compare the differences in content between journal papers and conference papers.Explore aspects such as publication trends, paper contributions, and thematic evolution to analyze the disparities in the presentation of information between journals and conference proceedings.Terahertz crystallography, terahertz optical materials, and terahertz optoelectronic radiation tend to have theoretical research outcomes published more frequently in journal papers. On the other hand, terahertz high-frequency communication and application systems, terahertz communication technology, terahertz detectors, terahertz imaging, and measurement technology lean towards technical and applied research, with a preference for publication in conference papers.The research findings of this study uncover differences in literature characteristics and research topic across journal papers and conference paperss. This contributes to a more nuanced interpretation of the patterns in basic research development, ultimately enhancing the accuracy of identifying disruptive technologies and other related aspects in future investigations.

Introduction

In the realm of cutting-edge foundational research, we find the epitome of scientific inquiry marked by disruptive innovation, interdisciplinary collaboration, and the juxtaposition of high risks and high rewards. Typically, outcomes in foundational research manifest in two primary forms: journal papers and conference papers. Journal papers often emphasize in-depth exploration of research questions, emphasizing the completeness, systematic nature, and scholarly qualities of proposed solutions. These articles cover a broad range of content, exhibiting enduring influence and dissemination effects within academic circles and the collective intellectual reservoir of humanity (Zhou, Y, 2013). On the other hand, conference papers tend to prioritize swift responses to research questions and inspire innovative thinking. Their focus lies in discovering new research directions and breakthroughs within the shortest possible time frame, proving invaluable for the rapid updating and expansion of knowledge structures in a particular field.

Small teams tend to prefer citing older and less influential papers, whereas larger teams are more inclined to cite the latest cutting-edge research, providing a new perspective for a comprehensive understanding of disruptive innovation (Wu et al., 2019). To further explore the evolutionary characteristics of the foundational research field in terms of publications, institutional entities, and research topics, this

paper employs a comprehensive approach incorporating bibliometrics, network analysis, and text mining methods. Firstly, it analyzes publication trends. Secondly, it utilizes organizational relationship network similarity to analyze the contribution of two types of publications. Thirdly, based on the evolution analysis of technical topics, it examines the distribution of different research topics across the two types of publications. Through this analysis, the paper reveals differences in literature characteristics and research topic across journal papers and conference paperss. This contributes to a more detailed interpretation of the patterns in basic research development and enhances the accuracy of identifying disruptive technologies and other related aspects in future investigations.

Literature review

The current status of research in the frontier areas of basic research

Since (Bush, 2020)proposed the linear development model from basic science to applied science and then to technological innovation, basic research has been the cornerstone of technological innovation for nearly a century. (Stokes, 2011) divided scientific research into four modes and proposed the dual-track model of basic science and technological innovation. (Narayanamurti & Odumosu, 2016) established the invention-discovery cycle model. (J. Chen et al., 2004) argued that the original innovation in basic research is the highest level of all innovations. In addition to discussing the concept of basic research and its relationship with innovation activities, researchers have also focused on exploring the theory of basic research. This includes the classification of basic research, the relationship between basic research and government support, and the activities of various entities involved in basic research. There has been less systematic discussion about the construction of the environment for basic research, as well as the interactive relationship between the subjects of basic research and the environment.

Over the years, bibliometrics and scientometrics methods have played a crucial role in the identification of frontier advances and strategic policy analysis in the field of basic research. Liu (2010), from the perspective of the distribution of scientific papers at the level of disciplines, countries, and institutions, analyzed the international cooperation patterns in basic research. Ma et al. (2015) based on data from the National Natural Science Foundation, proposed a comprehensive competitiveness index for basic research. They conducted a comparative analysis of the competitiveness of provincial-level regions in basic research in China and examined its changes over time. Chen et al. (2017), using three indicators - "activity index," "attraction index," and "efficiency index" - constructed a "comprehensive research capability index." They applied these indicators to compare and characterize the relative positions and competitive patterns of various countries in basic research in the field of science and technology. Zhang et al. (2018), based on the global overall research trends and representative research units, proposed indicators for the analysis of the competitive situation in basic research. The indicators covered strategic positioning, paper output, talent structure, and research patterns.

The current status of research on the differences in impact between journal papers and conference papers

Journals and conference papers, as two essential types of scientific literature, exhibit notable differences in publication cycles, document formats, and other aspects. Journal papers are publicly disseminated, and their academic levels, publication frequencies, and paper quantities are relatively stable. On the other hand, conference papers come in various publication formats, with significant variations in academic quality. However, in fields like computer science, communication, and others (such as IEEE top conferences), conference papers often demonstrate notable advancements and breakthroughs, drawing considerable attention from numerous peers in the respective fields. Given these distinctions in research outputs, scholars have extensively discussed the differences in influence between the two.

Some scholars argue that the academic impact of journal papers is higher than that of conference papers. Garvey (2014), in their analysis of the process of scientific literature production, considers conference papers as manuscripts for journal papers and suggests that the academic value of conference papers is lower than that of journal papers. Lisée et al. (2008), through bibliometric analysis of conference papers, find that the citation rate of conference papers is lower than that of journal papers. Wolek & Griffith (1974) point out that conference papers tend to be biased towards engineering and applied fields, suggesting their relatively lower "academic content" (Godin, 1998). Freyne et al. (2010), using the journal citation indicator from Web of Science as a measure, analyze and indicate that papers published in top conferences have a similar impact to those published in moderately ranked journals. Such studies mainly assert that scholars participate in academic conferences to share preliminary research results with peers, seeking feedback to refine subsequent research, ultimately leading to the successful publication of research outcomes in academic journals. Therefore, the impact of conference papers cannot be equated with that of journal papers.

Some scholars argue that the academic impact of conference papers is higher than that of journal papers, particularly in the field of computer science. Chen & Konstan (2010) point out that if a conference has a low acceptance rate, the citation frequency of papers published in that conference is similar to that of journal papers. Vrettas & Sanderson (2015) further indicate that the citation rate of papers from top computer science conferences is higher than that of journal papers, but the difference in citation rate between papers from mid to low-ranked conferences and journal papers is not significant. This kind of research primarily asserts that conference papers represent the final research outcomes and that conferences can replace certain engineering-related journal publications (Goodrum et al., 2001). It is suggested that there is no need for re-publication in journals, but this perspective is rooted in discussions among computer scientists and may not be universally applicable to other disciplinary areas.

Additionally, some scholars analyze the impact of the two types of literature from the perspective of the publication diffusion of journal and conference papers. Miguel-Dasit et al. (2006) suggest that journal papers originating from conferences are usually of high quality and more likely to receive high citations. In a study among

computer science scholars (Bar-Ilan, 2010), approximately 25-33% of CS-related conference papers were subsequently published in journals. Similar conversion rates from conferences to journals (30%) were reported in the field of computer vision publications (Eckmann et al., 2012). These conversion rates are lower than those in the medical field (Miguel-Dasit et al., 2007) but comparable to those in the field of information metrics (Aleixandre-Benavent et al., 2009). González-Albo & Bordons (2011) argue that the transition from conference to journal papers can be explained from the perspective of authors seeking to enhance research visibility and impact. Journal papers often have a greater potential to attract more citations than conference papers.

Domain Data and Research Framework

Domain Data

In recent years, terahertz technology, as a typical representative in the forefront of foundational research, operates at frequencies higher than microwaves and lower than infrared radiation. The energy levels fall between electrons and photons, giving it numerous unique properties compared to electromagnetic waves at other frequencies. In areas such as communication, radar, electronic warfare, electromagnetic weaponry, medical imaging, and security checks, terahertz technology holds tremendous potential for applications. It has been recognized by the United States as one of the 'Top Ten Technologies Changing the Future World.' Currently, terahertz technology is gaining increasing attention worldwide due to its distinctive capabilities and broad prospects. It is internationally acknowledged as a contested area in high-tech fields, and its research and applications are considered to have significant strategic implications for future warfare and national security (Qian, 2022).

Based on this, the present study utilizes the Web of Science (WoS) Core Collection database as the data source, using terahertz field papers as an example to conduct a comparative analysis of the differences between journal papers and conference papers in the field of basic research. The literature search formula is TS=("terahertz*" OR "terahertz*"), refining the Web of Science index to SCI-E, ESCI; refining the publisher to IEEE; the time range is from January 1, 2004, to December 31, 2023, with the search conducted on June 9, 2023. The results were downloaded in "plain text" format, and after removing duplicates, a total of 44,683 papers were obtained, including 33,057 journal papers and 11,626 conference papers.

Research Framework

This study, depicted in Figure 1 as the main research framework, analyses the differences between journal papers and conference papers at the forefront of basic research from three perspectives: publication trends, paper contributions, and thematic evolution.

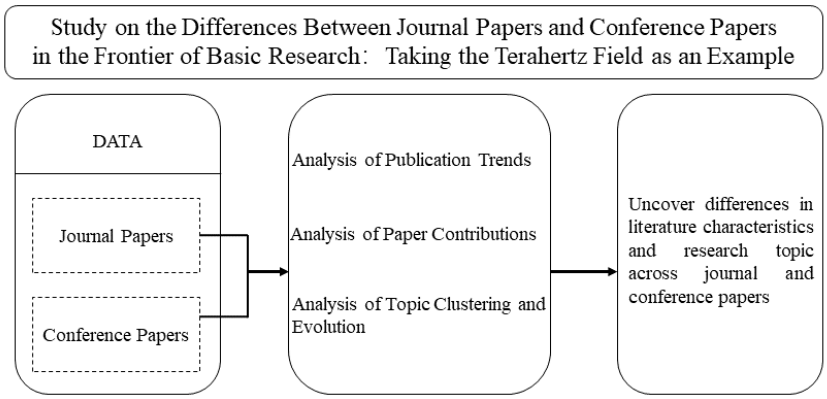


Figure 1. Research framework diagram.

Analysis Methodology for Paper Contributions

To effectively analyze the contributions of journal and conference paper collections to the total paper collection, this paper adopts an analysis approach from the perspective of the similarity of institutional relationship networks (co-occurrence network, citation network). This approach aims to present the comparison results of paper contributions more comprehensively and accurately. The coupling of institutional relationship networks is manifested in the coupling of nodes and structural coupling between networks. Node coupling in institutional relationship networks reveals the institutional associations formed by the correspondence between network nodes, while structural coupling arises from the consistency of network edges. To measure the similarity between institutional relationship networks, we propose a new method that evaluates their node coupling strength and structural coupling strength. The node coupling strength of institutional relationship networks is calculated based on the similarity of PageRank values of coupled nodes in the two networks, while the structural coupling strength is calculated based on the similarity of Weight values of coupled edges in the two networks.

Methodology for Theme Identification

Based on the overall paper collection in the terahertz field, topic clustering is performed using the titles and abstracts of the papers. The data undergoes preprocessing, including standardizing the case, removing punctuation, part-of-speech tagging, lemmatization, and eliminating stop words. Subsequently, the optimal number of topics is determined based on the topic perplexity method and LDAvis (Blei et al., 2003). The matching probability values between papers and topics are then exported to identify the topics to which different papers belong, and names are assigned to different topics. Finally, based on the identification results, the literature volume of different topics over time is analyzed to further conduct a topic evolution analysis.

Empirical Analysis

Analysis of Publication Trends

The publication trend is illustrated in Figure 2. From 2004 to 2023, the global terahertz field showed a rapid growth in the number of publications, with a noticeable increase in 2013, surpassing 2300 papers globally. Journal papers also exhibit a growing trend, closely aligning with the overall publication trend. However, the growth trend of conference papers is not as pronounced and consistently remains below 1000 papers. Particularly around 2020, journal papers continue to show growth, while conference paper publications experience a declining trend.

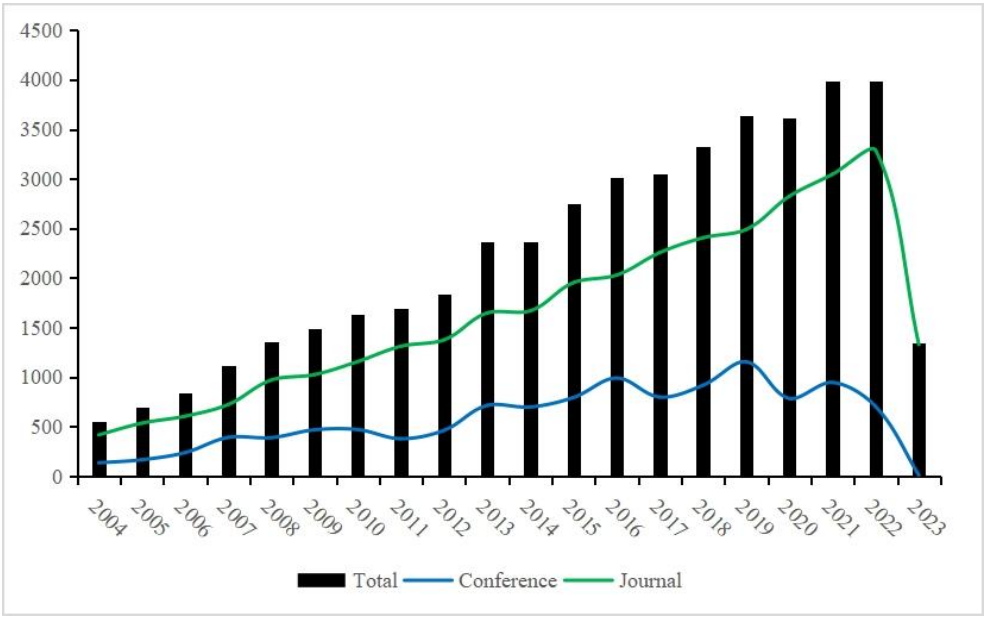


Figure 2. Publication trends of terahertz field journal papers and conference papers.

Figure 3 presents the publication output of the top 10 countries in the field of terahertz research. China has published 14,080 papers, while the United States has published 8,317 papers, indicating that China significantly surpasses the United States. In 2004, the United States had the highest publication output, with China ranked 6th. However, China's publication output started to increase annually and surpassed the United States in 2014, becoming the world's leading country in terms of terahertz research publications. Since 2014, China has consistently maintained its position as the top contributor with a significantly higher publication output than the United States.

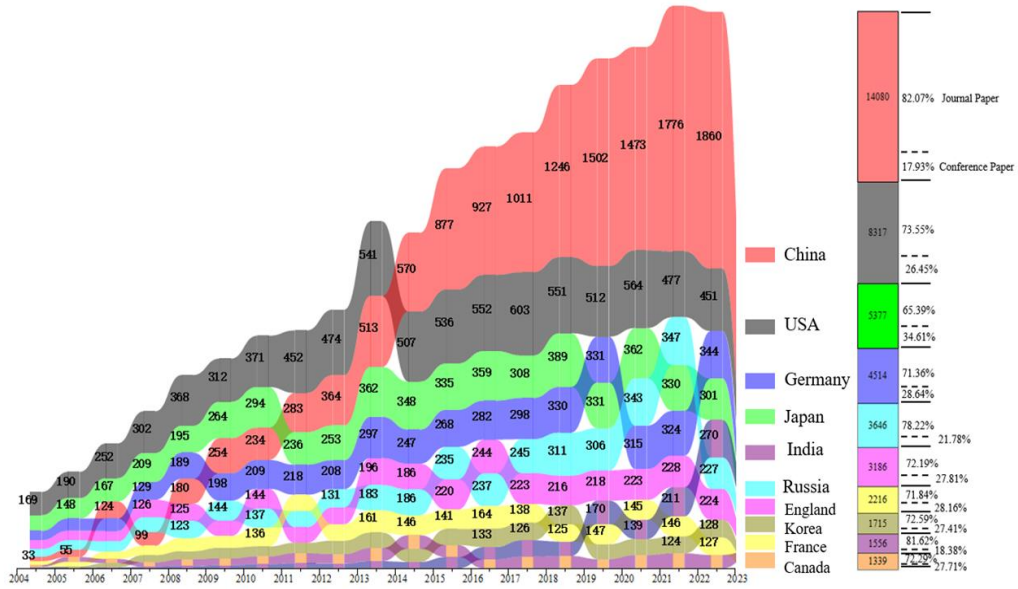


Figure 3. Trend chart of publication output in the top 10 countries in terms of publication quantity.

Although China and the United States have a similar output in conference papers, conference papers only account for 17.93% of the total. It is worth mentioning that among the top 10 countries with the highest publication output, China, India, and Russia, as developing countries, have a higher proportion of journal papers compared to the other seven developed countries. China and other developing countries tend to publish more journal papers in the field of terahertz research compared to countries like the United States.

Comparison of Contributions between Journal Papers and Conference Papers

This study conducted separate analyses for the overall period and sub-periods, namely 2004-2008, 2009-2013, 2014-2018, and 2019-2023. The analyses focused on the co-occurrence network and citation network among institutions in the overall dataset, journal paper dataset, and conference paper dataset. Tables 1 and 2 present the findings.

Based on the co-occurrence network, it was observed that the number of nodes and edges in the journal paper dataset is very close to that of the overall dataset, while the number of nodes and edges in the conference paper dataset is relatively smaller. This indicates that institutional collaborations in the field of terahertz research are predominantly reflected in journal papers, whereas collaborations in conference papers are relatively weaker. This pattern remained consistent across the four time periods.

Additionally, the similarity between the co-occurrence network of journal papers and the overall dataset was found to be 0.98365, which is very close to 1. On the other hand, the similarity between the co-occurrence network of conference papers and the overall dataset was 0.75777, lower than the similarity value for journal papers. This

similar pattern was also observed across the four time periods. These findings suggest that journal papers exhibit a higher similarity to the overall dataset in terms of institutional collaborations in the co-occurrence network. Journal papers make a greater contribution and hold higher value compared to conference papers in the co-occurrence network of institutions.

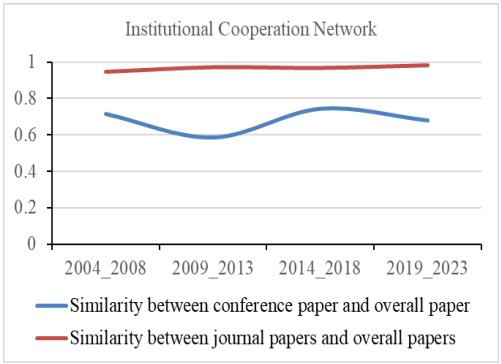
Table 1. Overview and similarity of co-occurrence networks in institutions.

Net	Style	All Time	2004-2008	2009-2013	2014-2018	2019-2023
Total dataset co-occurrence network in institutions	Number of nodes	10197	1502	2862	4529	6053
	Number of edges	62685	4129	14047	22901	31959
Journal paper collection co-occurrence network	Number of nodes	8883	1276	2441	3875	5500
	Number of edges	57247	3520	12604	20478	29833
	Similarity to the total dataset	0.98365	0.94663	0.97305	0.96915	0.98464
Conference paper collection co-occurrence network	Number of nodes	3522	554	1057	1791	1733
	Number of edges	10626	952	2342	4522	4103
	Similarity to the total dataset	0.75777	0.71610	0.58543	0.74481	0.68040

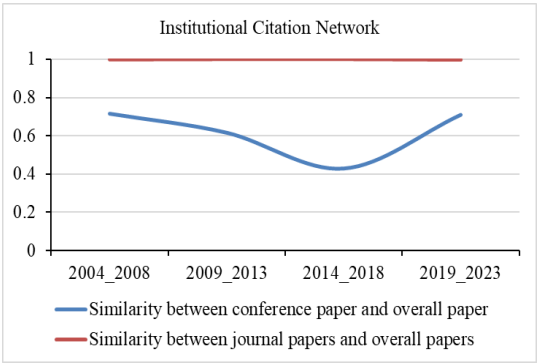
Table 2. Overview and similarity of citation networks in institutions.

Net	Style	All Time	2004-2008	2009-2013	2014-2018	2019-2023
Total dataset citation network in institutions	Number of nodes	9150	1176	2863	5096	8212
	Number of edges	533257	14158	74550	195149	383699
Journal paper collection citation network	Number of nodes	8923	1139	2846	5087	8003
	Number of edges	527351	13920	74272	194908	377791
	Similarity to the total dataset	0.99910	0.99799	0.99987	0.99993	0.99663
Conference paper collection citation network	Number of nodes	2758	241	289	273	2585
	Number of edges	19813	656	793	684	18043
	Similarity to the total dataset	0.71876	0.71660	0.61693	0.42938	0.71088

Based on the four designated periods, the comparison of the similarity trends between journal papers and conference papers in the co-occurrence networks with the total dataset is presented. Additionally, the similarity trends of journal papers and conference papers in the institutional citation networks with the total dataset are also compared, as shown in Figure 4. In both institutional collaboration networks and institutional citation networks, the similarity of journal papers to the overall papers remains close to 1, consistently higher than the similarity of conference papers to the overall papers. This indicates that, whether in institutional collaboration networks or institutional citation networks, journal papers contribute more significantly compared to conference papers. The similarity of conference papers to the overall papers in institutional collaboration networks fluctuates but consistently maintains around 0.7. On the other hand, the similarity of conference papers to the overall papers in institutional citation networks experienced a significant decline from 2014 to 2018, suggesting that the contribution of conference papers to institutional citation networks is relatively weaker compared to institutional collaboration networks.



The comparative trend of similarity between journal papers and conference papers in institutional co-occurrence networks



The comparative trend of similarity between journal papers and conference papers in institutional citation networks

Figure 4. The similarity changes in institutional collaboration networks and institutional citation networks.

Topic Clustering and Evolution Analysis

Analysis Methodology for Paper Contributions

Calculate the perplexity with the change in the number of topics, as shown in Figure 5 (Left). Identify the inflection point as the optimal number of topics, which is 8. From Figure 8 (Right), it can be observed that the distribution of each topic is sparse with fewer crossovers, indicating a good result in topic identification. The results of topic clustering are shown in Table 3.

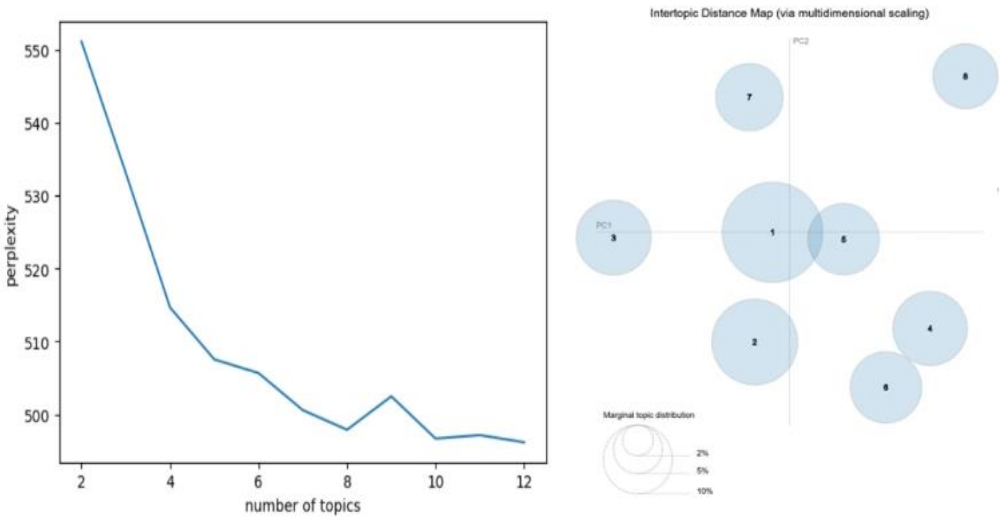


Figure 5. Selection of a number of topics.

Table 3. Overall dataset topic clustering results in the Terahertz domain.

Topic	Topic Concepts	Total number	Journal	Conference
Topic 1: Terahertz detectors	Emphasis is placed on terahertz detector technologies, including detectors based on GaAs materials, photodetectors, and detectors operating in room temperature and resonance modes.	5204	3445(66%)	1759(34%)
Topic 2: Terahertz crystallography	Focus on crystallographic research in the terahertz range, including studying crystal structures, domain distribution, characteristics of solid materials, and phase transitions at different temperatures.	3378	2852(84%)	526(16%)
Topic 3: Terahertz communication technology	Main focus on communication technology in the terahertz range, including terahertz antenna technology, waveguide technology, filter technology, and integrated circuit design.	4676	3076(66%)	1600(34%)
Topic 4: Terahertz optical materials	Research optical materials in the terahertz range, including materials based on graphene, surface plasmon resonance, multilayer structures, etc. These materials are used for applications such as absorption, transmission, modulation, and sensing of terahertz waves.	9108	7503(82%)	1605(18%)
Topic 5: Terahertz spectroscopy	Focus on spectroscopic techniques in the terahertz range, including using terahertz waves for time-domain spectroscopic analysis, measurement of refractive index and absorption coefficient of materials, research in the field of thin films, exploration of dynamic processes, and analysis of transmission and scattering characteristics.	3393	2588(76%)	805(24%)
Topic 6: Terahertz optoelectronic radiation	Focus on radiation phenomena and generation mechanisms in the terahertz range, including radiation from terahertz fields, wave generation, nonlinear optical effects, pulse excitation, and interactions between terahertz light and electrons.	9562	7726(81%)	1836(19%)
Topic 7: Terahertz high-frequency communication and application systems	Focus on communication system technologies in the terahertz range, including high-frequency and high-bandwidth communication systems, signal modulation, transmission technologies, and exploring applications of terahertz waves in the spatial domain.	3800	2114(56%)	1686(44%)
Topic 8: Terahertz imaging and measurement technology	Focus on imaging and measurement technologies in the terahertz range, including imaging methods, resolution improvement, as well as experimental validation of detection and measurement methods, measurement accuracy, and sensitivity.	5562	3753(67%)	1809(33%)

It can be observed that the terahertz domain is divided into 8 research topics. Among them, Topic 6 and Topic 4 have a relatively high total number of papers, with 9562 and 9108 papers respectively, while the number of papers on other topics is around 4000. In terms of paper type distribution, journal papers are more prevalent in Topics 2, 4, and 6, accounting for over 80%, indicating a preference for publishing theoretical research results in journal papers. On the other hand, conference papers have a larger proportion in Topics 7, 3, 1, and 8, all exceeding 30%, especially in Terahertz High-Frequency Communication and Application Systems, with a proportion of 44.37%, far exceeding the average conference paper proportion of 26.02%. This suggests that achievements in technology and applied research are more inclined to be published in conference papers.

Topic Evolution Analysis

Based on the results of topic identification, the literature volume of different topics over time in various periods is statistically analyzed, as shown in Figure 6. Topics 1, 2, 5, and 6, while showing a slight increase in the number of publications in each period, generally maintain a relatively constant state. On the other hand, Topics 3, 4, 7, and 8 exhibit an overall significant growth trend in literature volume, indicating

an increasing attention to their research in recent years, with the output growing annually.

Moreover, in terms of absolute publication volume, Topic 6 has consistently received high attention in each period, maintaining a consistently high publication output and being one of the research hotspots. At the same time, Topic 4 shows a clear evolutionary growth trend with a substantial increase in publications, garnering increasing attention in recent years and becoming one of the most prominent research hotspots. Topics 2 and 5, with consistently lower publication volumes, have lower levels of attention. Although Topics 3 and 7 generally have a lower overall attention level, their publication volumes have been growing in recent years, suggesting the potential to become new technological research hotspots in the future.

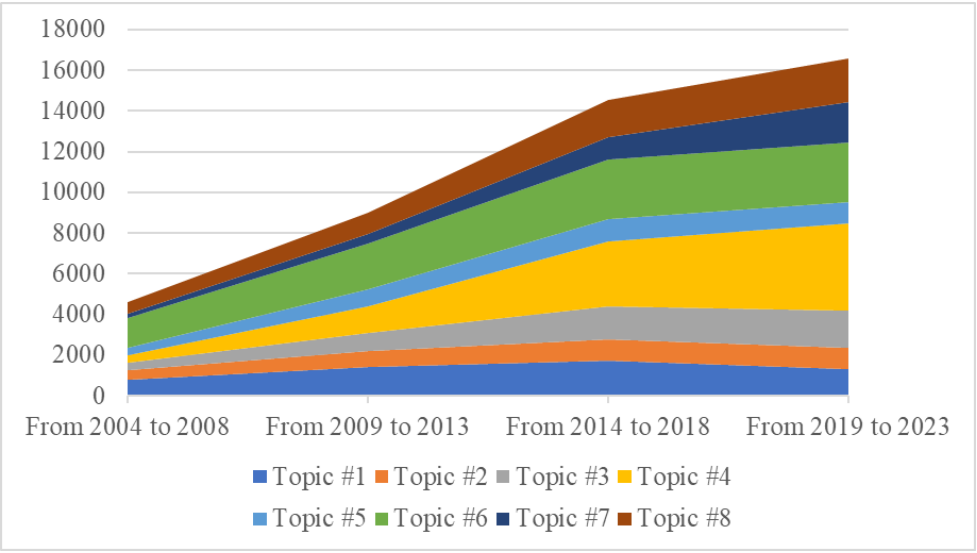


Figure 6. The evolutionary trend of the total paper collection's topic literature volume over time.

Discussion and Outlook

With the evolution of the international science and technology competitive landscape, the pursuit of cutting-edge basic research has become a focal point for major technological powers. This trend places higher demands on the innovative application of literature and information methods. To more accurately carry out the identification of disruptive technologies and frontier hotspots in the field of basic research, and to explore from a richer perspective, this study investigates the differences in published literature between journal papers and conference papers. Taking the terahertz domain as an example, the research empirically analyses the differences in publication trends, paper contributions, and topic evolution between the two types of papers. This exploration aims to further discuss the characteristics and patterns reflected in basic research papers on different types of papers .

The research results indicate that:

Over the past 20 years, the global publication volume in the terahertz domain has shown an overall increasing trend. The top 10 countries with the highest publication

volumes in the terahertz domain are China, the United States, Japan, Germany, Russia, the United Kingdom, France, South Korea, India, and Canada. China, in particular, has a significantly higher publication volume in the terahertz domain compared to other countries. Among the top 10 countries with the highest publication volumes, the proportion of journal papers from developing countries such as China, India, and Russia is higher than that of the other seven developed countries. Developing countries, including China, tend to publish more journal papers in the terahertz domain compared to developed countries like the United States.

In terms of the contributions of the two types of literature, whether in institutional citation networks or institutional co-occurrence networks, the similarity of journal papers to the overall paper collection is higher than that of conference papers. Considering that the quantity of journal papers is generally higher than that of conference papers, the domain contribution value of journal papers remains higher. Over time, the similarity of journal papers to the overall papers in both institutional co-occurrence networks and institutional citation networks is higher than the similarity of conference papers to the overall papers. In most domains, institutional collaboration is more prominent in journal papers, while institutional collaboration in conference papers is relatively weaker.

Regarding topic identification, the terahertz domain comprises 8 research topics: terahertz detectors, terahertz crystallography, terahertz communication technology, terahertz optical materials, terahertz spectroscopy, terahertz optoelectronic radiation, terahertz high-frequency communication and application systems, and terahertz imaging and measurement technology. Results inclined towards theoretical research preferentially appear in journal papers, while results leaning towards technology and applied research are more likely to be published in conference papers. Currently, terahertz optoelectronic radiation and terahertz optical materials are two major research hotspots, while terahertz communication technology and terahertz high-frequency communication and application systems are expected to become new research hotspots in the future, garnering higher research attention.

In the application of traditional literature and information methods to the identification of cutting-edge and disruptive research in basic research, there has been a greater focus on the content of large-scale datasets in the field and their internal relationships. To some extent, this approach has overlooked the mutual relationship between the characteristics of literature across different mediums and the evolving trends of research subjects. While this study has certain limitations in terms of data scale, field selection, and method choice, the research results reveal differences in the characteristics of literature across different mediums and research topics. This insight can contribute to a more detailed interpretation of the development patterns in basic research and enhance the accuracy of identifying disruptive technologies in the future.

Acknowledgments

The study is supported by Youth Innovation Promotion Association Funding Projects (2023183) of Chinese Academy of Sciences, and Sichuan Provincial Science and Technology Plan Soft Science Project: “Innovative Research on the Operation

Mechanism of National Laboratories and Tianfu Laboratories" (Grant No.2023JDR0013).

References

Aleixandre-Benavent, R., González-Alcaide, G., Miguel-Dasit, A., Navarro-Molina, C., & Valderrama-Zurián, J. (2009). Full-text publications in peer-reviewed journals derived from presentations at three ISSI conferences. *Scientometrics*, *80*(2), 407–418.

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. *Scientometrics*, *83*(3), 809–824.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bush, V. (2020). *Science, the endless frontier*. Princeton University Press.

Chen, J., & Konstan, J. A. (2010). Conference paper selectivity and impact. *Communications of the ACM*, *53*(6), 79–83.

Chen, J., Song, J., Ge, C., & Zhu, X. (2004). On the basic research and original innovation. *Studies in Science of Science*, *22*(03), 317–321.

Chen, K., Zhang, Y., & Mu, R. (2017). The international comparison of the basic research capability in the science & technology field————Evidence from the field of energy storage. In *Studies in Science of Science* (CNKI; Vol. 35, Issue 01, pp. 34–44).

Eckmann, M., Rocha, A., & Wainer, J. (2012). Relationship between high-quality journals and conferences in computer vision. *Scientometrics*, *90*(2), 617–630.

Freyne, J., Coyle, L., Smyth, B., & Cunningham, P. (2010). Relative status of journal and conference publications in computer science. *Communications of the ACM*, *53*(11), 124–132.

Garvey, W. D. (2014). *Communication: The essence of science: Facilitating information exchange among librarians, scientists, engineers and students*. Elsevier.

Godin, B. (1998). Measuring knowledge flows between countries: The use of scientific meeting data. *Scientometrics*, *42*, 313–323.

González-Albo, B., & Bordons, M. (2011). Articles vs. Proceedings papers: Do they differ in research relevance and impact? A case study in the Library and Information Science field. *Journal of Informetrics*, *5*(3), 369–381.

Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing & Management*, *37*(5), 661–675.

Lisée, C., Larivière, V., & Archambault, É. (2008). Conference proceedings as a source of scientific information: A bibliometric analysis. *Journal of the American Society for Information Science and Technology*, *59*(11), 1776–1784.

Liu, Y. (2010). Analysis of the international collaboration of chinese basic research based on bibliometrics. In *Forum on Science and Technology in China* (CNKI; Issue 03, pp. 149–155).

Ma, Y., Wang, G., & Wan, Y. (2015). Comparative analysis on the regional medical science basic research in china based on NSFC. *Science and Technology Management Research*, *35*(17), 71–76.

Miguel-Dasit, A., Martí-Bonmatí, L., Aleixandre, R., Sanfeliu, P., & Valderrama, J. C. (2006). Publications resulting from Spanish radiology meeting abstracts: Which, Where and Who. *Scientometrics*, *66*(3), 467–480.

Miguel-Dasit, A., Martí-Bonmatí, L., Sanfeliu-Montoro, A., Aleixandre, R., & Valderrama, J. C. (2007). Scientific papers presented at the European Congress of Radiology: A two-year comparison. *European Radiology*, *17*, 1372–1376.

Narayanamurti, V., & Odumosu, T. (2016). *Cycles of invention and discovery: Rethinking the endless frontier*. Harvard University Press.

Qian, Y. (2022). Terahertz: The next generation disruptive foundational technology. *Hangzhou Science & Technology, 53*(04), 23–26.

Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.

Vrettas, G., & Sanderson, M. (2015). Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology*, *66*(12), 2674–2684.

Wolek, F. W., & Griffith, B. C. (1974). Policy and informal communications in applied science and technology. *Science Studies*, *4*(4), 411–420.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382.

Zhang, Z., Chen, Y., Tao, C., Xu, J., & Tian, Q. (2018). Bibliometric analysis on international competitive situation of quantum information research. In *World Sci-Tech R & D* (CNKI; Vol. 40, Issue 01, pp. 37–49).

Zhou, Y. (2013). A comparative analysis between periodical paper and proceedings literature: A case of computer science software engineering. *Documentation, Information & Knowledge, 0*(6), 114.

https://doi.org/10.51408/issi2025_087

Synergy Between Science And Technology In University-Industry Innovation Ecosystems: A Cross-National Comparison Of Elite Academic Partnerships In China, Germany, And The United States

Hui Zhang[1], Hui Fu[2], Ying Huang[3]

[1] *hui_zhang@whu.edu.cn*, [2] *2019301040194@whu.edu.cn*

Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)

School of Information Management, Wuhan University, Wuhan (China)

[3] *ying.huang@whu.edu.cn*

Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)

School of Information Management, Wuhan University, Wuhan (China)

Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)

Abstract

University-industry collaboration serves as a critical driver of technological innovation, significantly contributing to national economic growth and enhancement of global competitiveness. This study addresses the pivotal challenge of optimizing such partnerships and improving the commercialization efficiency of scientific breakthroughs through an empirical investigation of 26 elite universities from China's C9 League, Germany's Universities of Technology Alliance, and the United States Ivy League (2000-2020). Grounded in the knowledge spiral framework, the research employs integrated bibliometric analysis and social network mapping to systematically compare cross-national innovation ecosystems. Findings indicate that while German and American institutions demonstrate superior performance in knowledge co-creation dynamics, Chinese universities lead in patent authorization volume yet face challenges in university-industry collaboration rates and commercialization outcomes. Network analysis reveals distinct structural patterns: Chinese co-authorship networks exhibit institutional dominance with limited enterprise engagement, whereas patent collaboration forms university-centric clusters maintaining exclusive enterprise partnerships. These insights provide actionable pathways for enhancing knowledge transfer mechanisms and inform evidence-based policy formulation in national innovation systems.

Introduction

With the evolution of the new round of scientific and technological revolution, technological innovation has increasingly become an important means for countries to promote economic development and enhance competitiveness. At the same time, scientific research has shown the characteristics of interdisciplinarity and comprehensiveness. Significant breakthroughs in scientific research rely

increasingly on interdisciplinary, cross-domain, cross-institutional, and cross-national cooperation. Research cooperation has become a significant trend in global scientific research progress. To promote technological innovation and research cooperation, governments worldwide have placed scientific and technological innovation at the core of national development and promulgated policies to enhance national scientific and technological innovation capabilities. In the 1980s, the United States introduced the Bayh-Dole Act (Kenney & Patton, 2009) to address the problem of idle research achievements and reduce economic competition pressure, encouraging universities and enterprises to cooperate in research projects and promoting technological innovation and technology transfer. Germany has always attached great importance to technological innovation and formulated strategies such as the High-Tech Strategy to provide policy guidance for the cooperation between universities and enterprises. It also builds innovation clusters and platforms to construct an innovation network and promote interdisciplinary cooperation. In recent years, China has increasingly emphasized the transformation of scientific and technological achievements and university-industry cooperation(P. s. R. o. China, 2021, 2022). Universities and enterprises around the globe are proactively exploring and implementing innovative cooperation patterns under the guidance of established policies.

As an important driver of technological innovation, universities are regarded as an important source of new knowledge for enterprises (Rast, Khabiri, & Senin, 2012). Universities serve as knowledge producers and guides, supplying enterprises with the latest theories and insights. By absorbing diverse knowledge from universities and offering technical support, organizations facilitate the transformation of research outcomes into practical applications. Consequently, the "university-enterprise" cooperation pattern has emerged as a crucial method for universities to produce, utilize, and transform knowledge within the framework of open innovation. Internationally, leading universities in Germany and the United States boast exceptional research talent and facilities, forming strong partnerships with local businesses. The foundation of university-industry cooperation in Germany stems from the "dual system" of vocational and technical education, which has significantly enhanced collaboration among industry, academia, and research institutions and the application of scientific research findings (Xiao, 2016). The United States, as the birthplace of industry-academia-research teaching, has received substantial government support for university-industry cooperation (Foundation, 2018). Universities actively explore and practice university-industry cooperation patterns, from joint research to company incubators, forming various university-industry cooperation paths.

Currently, relevant research on scientific and technological innovation cooperation between universities and enterprises at home and abroad mainly focuses on cooperation patterns, cooperation performance evaluation, cooperation network

evolution, and technology transfer. X. Wang, Wang, and Liu (2005) proposed six cooperation patterns based on league forms and participating entities. Ding, Huang, and Guo (2010), based on the practice of university-industry cooperation in higher vocational colleges, proposed university-industry cooperation patterns led by enterprises and universities respectively. Kwon, Park, So, and Leydesdorff (2012) based on the triple helix theory, constructed innovation indicators to analyze the structural pattern of Korean universities' participation in university-enterprise cooperation. S. Wang (2020) constructed a pattern for evaluating the technological innovation performance of universities. F. Liu, Ma, and Jiang (2011) studied the evolution path of the industry-university-research cooperation network based on "985 universities" from patent cooperation data. Dang, Jasovska, Rammal, and Schlenker (2019) analyzed the knowledge transfer between university-enterprise cooperation by studying the university-industry cooperation methods of ten Australian universities. Scholars' research on university-industry cooperation is mainly based on "patent" data. The sample universities in the research generally focus on specific regions (such as the Yangtze River Delta region in China) or specific fields (such as Australian business schools). The data on innovation cooperation achievements lack diversity, and there are few industry comparisons among international top universities. As one of the important forms of the achievements of university-enterprise innovation cooperation, co-authored articles of industry and academia are also an important indicator reflecting the characteristics of university-enterprise cooperation (Jianjie Guo, Xie, Wang, & Wang, 2019).

Existing literature predominantly examines the scientific (articles) and technological (patents) dimensions in isolation, with limited focus on their synergistic relationship. Additionally, there is a lack of research on university-industry collaboration in top universities across different countries. As leading academic institutions in China, Germany, and the United States, the C9 Alliance, TU9, and Ivy League universities have significant domestic and international influence. These universities are well-established in research mechanisms and highly active in industry collaborations, and university-industry collaboration models in these institutions are highly representative. Therefore, this study aims to combine university paper data and patent data to analyze the state of scientific innovation cooperation between university alliances and industry in China, Germany, and the United States from an international perspective. The findings will offer valuable insights to promote university-industry cooperation and accelerate the technological innovation process.

Conceptual Model and Framework

Conceptual Model

The cooperation between universities and industries primarily revolves around the transfer of knowledge. In this process, both universities and companies invest

various resources, including scientific research personnel, research facilities, funding, technical support, and diverse knowledge. The goal is to create new knowledge and achieve innovative results, such as enhancing the value of knowledge, fostering scientific and technological advancements, and developing talent. Numerous scholars have examined this process from different angles and have proposed various theoretical models to explain it. The "Triple Helix Model" proposed by Etzkowitz and Leydesdorff emphasizes the important roles of universities, enterprises, and government in the process of knowledge production and dissemination (Etzkowitz & Leydesdorff, 1996). Hu, Zhu, and Ma (2011) systematically analyzed the interrelated constraints among various factors in university-industry-research cooperation and constructed a system dynamics model of university-industry-research cooperation. R. Wu, Liu, and Li (2021) combined the SECI theory in knowledge management theory and the knowledge collaborative innovation mechanism to construct a SECI (Socialization, Externalization, Combination, Internalization) knowledge transfer model based on the "collaborative pool" to reveal the knowledge transfer phenomenon in the process of university-enterprise cooperation. The SECI knowledge spiral theory proposed by Nonaka and Takeuchi in 1994 (Nonaka, 1994), is considered one of the most classic theoretical models in the field of knowledge transfer. They believe that knowledge creation is essentially a continuous transformation, recombination, and utilization process of tacit and explicit knowledge. Tacit knowledge includes untextualized experiences such as thinking patterns and intuition, while explicit knowledge refers to knowledge that can be textualized and disseminated. The SECI model believes that the process of knowledge transfer includes four stages: socialization, externalization, combination, and internalization. Explicit and tacit knowledge interact and transform in different stages, forming a virtuous knowledge creation cycle. The SECI model can systematically summarize the knowledge flow pattern between universities and enterprises and provide a theoretical basis for understanding the knowledge creation process in university-industry cooperation. Therefore, this study introduces the SECI knowledge spiral theory and combines the input-output elements in university-enterprise cooperation to construct a university-industry cooperation model based on the SECI knowledge spiral theory, as shown in Figure 1. In this model, universities and enterprises contribute resources that facilitate the interaction and transformation of their diverse knowledge, leading to knowledge creation and innovative outcomes. These innovative achievements can be represented by both quantifiable elements, such as the number of co-authored articles and cooperative patents, as well as non-quantifiable elements, including talent development and institutional competitiveness.

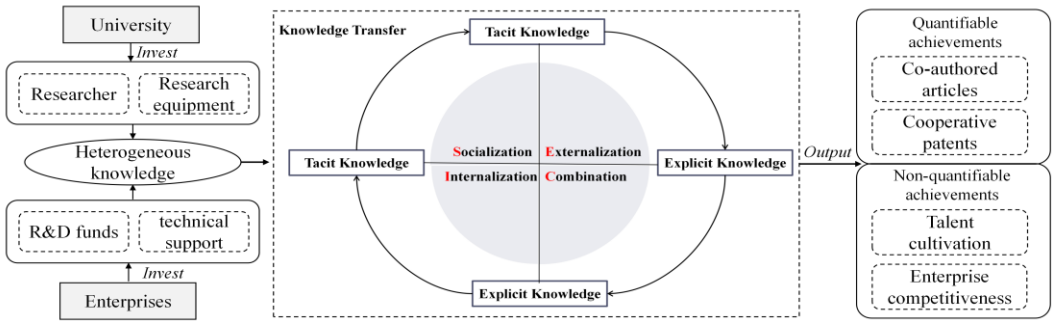


Figure 1. University-Industry Cooperation Model based on the Knowledge Spiral Theory.

Research Framework

This study commences from the article and patent data of universities and explores the characteristics and patterns of university-industry cooperation among different universities in China, Germany, and the United States through bibliometric and cooperation network analysis methods. Figure 2 shows the overall framework of this study. The study is divided into three sections: data collection and processing, bibliometric analysis, and social network analysis. Figure 2 shows the overall framework of this study. The bibliometric analysis focuses on article and patent data, comparing the proportion and temporal trends of university-industry collaboration in articles and patents to analyze the collaborative models and evolution of top universities in different countries. The social network analysis, on the other hand, examines the collaboration networks of universities in China, Germany, and the United States, based on articles and patents, to explore the structure, strength, and pathways of cooperation between universities and industry, providing insights into the distinct advantages and characteristics of university-industry collaborations across the three countries.

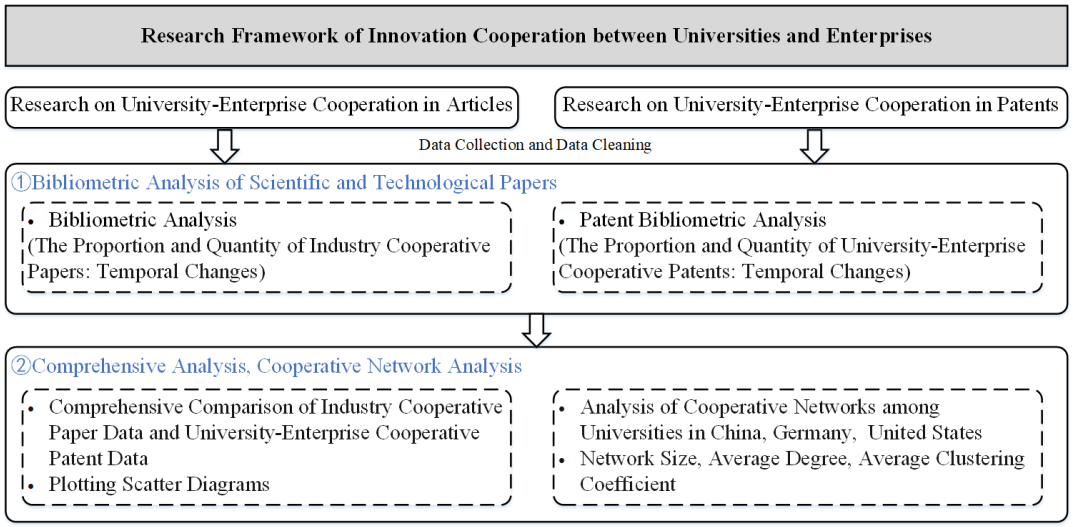


Figure 2. Research Framework of University-Enterprise Innovation Cooperation.

In terms of data collection and processing, this study intends to select the university-industry cooperation data of top universities in China, Germany, and the United States as samples for analysis. We have selected a total of 26 universities from China's C9 League, Germany's Universities of Technology (TU9), and the US Ivy League to research university-industry cooperation. This study aims to gain insights into the collaboration situations of top universities in each country. As leading institutions in China, Germany, and the United States, the C9 League, TU9, and Ivy League hold significant influence both nationally and globally. These universities have established robust scientific research mechanisms and actively engage in cooperation with enterprises. Analyzing the current state and patterns of university-industry cooperation in these institutions will provide valuable insights. We select the Incites and Web of Science databases to obtain university-industry cooperation data and the Derwent Innovations to obtain university patent data. Derwent Innovations is one of the world's most comprehensive patent information databases, providing unique patent indexing, which is helpful for studying the patent data of universities in various countries. Since the patent examination process generally takes 18 months after application, to ensure the accuracy and consistency of the data, we limit the retrieval time range of articles and patents to be unified between 2000 and 2020. There is no mark in the patent data indicating whether there is university-enterprise cooperation. In this study, "university-enterprise cooperation patents" are defined as patents jointly researched and applied by universities and enterprises, where universities and enterprises are in a partnership relationship, and the judgment basis is that both university and enterprise types are included in the patent applicant

field (X. Wang et al., 2005). After conducting a search based on specific terms, a total of 88,481 articles on industry cooperation were obtained. After excluding missing values and outliers and performing deduplication, we were left with 61,049 articles. For the patent data, we carried out cleaning, word segmentation, and filtering. We retained only those patents that listed both university and enterprise applicants, resulting in a final total of 15,892 patent entries.

Results

Quantity and Temporal Variation in Co-authored Articles

By analyzing the industry cooperation article data of universities, it is found that American and German universities perform well in co-authoring articles with enterprises. Harvard University has the highest number of industry co-authored articles, and Princeton University has the highest proportion of industry co-authored articles, more than twice the proportion of Tsinghua University's industry articles. Table 1 shows the numbers of the industry-cooperation articles of the sample universities from 2000 to 2020.

Table 1. The numbes of Industry-Cooperation Articles of the Sample Universities (2000-2020).

Country	University	Industry Collaboration Articles Count	Industry Collaboration Proportion
China	Tsinghua University	5571	2.79%
	Shanghai Jiao Tong University	4390	2.28%
	Peking University	3541	1.97%
	Zhejiang University	3296	1.71%
	Fudan University	2373	1.90%
	Xi'an Jiaotong University	2296	2.13%
	University of Science and Technology of China	2029	1.73%
	Harbin Institute of Technology	1480	1.31%
	Nanjing University	1257	1.34%
Germany	Technical University of Munich	6177	5.03%
	RWTH Aachen University	4436	5.70%
	Dresden University of Technology	3921	4.05%
	Karlsruhe Institute of Technology	2369	5.08%
	Technical University of Berlin	996	4.08%

	University of Stuttgart	906	4.20%
	Darmstadt University of Technology	874	3.85%
	University of Hanover	664	3.71%
	Brunswick Technical University	619	4.35%
United States	Harvard University	16874	4.87%
	University of Pennsylvania	5792	4.49%
	Columbia University	5350	4.47%
	Cornell University	4134	3.92%
	Princeton University	3305	6.61%
	Yale University	3206	3.11%
	Brown University	1434	3.17%
	Dartmouth College	1153	4.18%

The data in the table indicates that the proportions of industryly co-authored articles from universities in Germany and the United States are generally higher. Seventeen universities have proportions exceeding 3.5%, suggesting that the top institutions in these countries are more active in collaborating with enterprises for co-authorship. In contrast, while the number of industryly co-authored articles from Chinese universities is comparable to that of Germany, the proportion remains low. Only Tsinghua University and Shanghai Jiao Tong University have proportions of industryly co-authored articles that exceed 2%. This highlights a significant opportunity for improvement in collaboration between Chinese universities and enterprises.

Figures 3 and 4 illustrate the trends in the number and proportion of industryly co-authored articles for different countries and universities, analyzed by time and university.

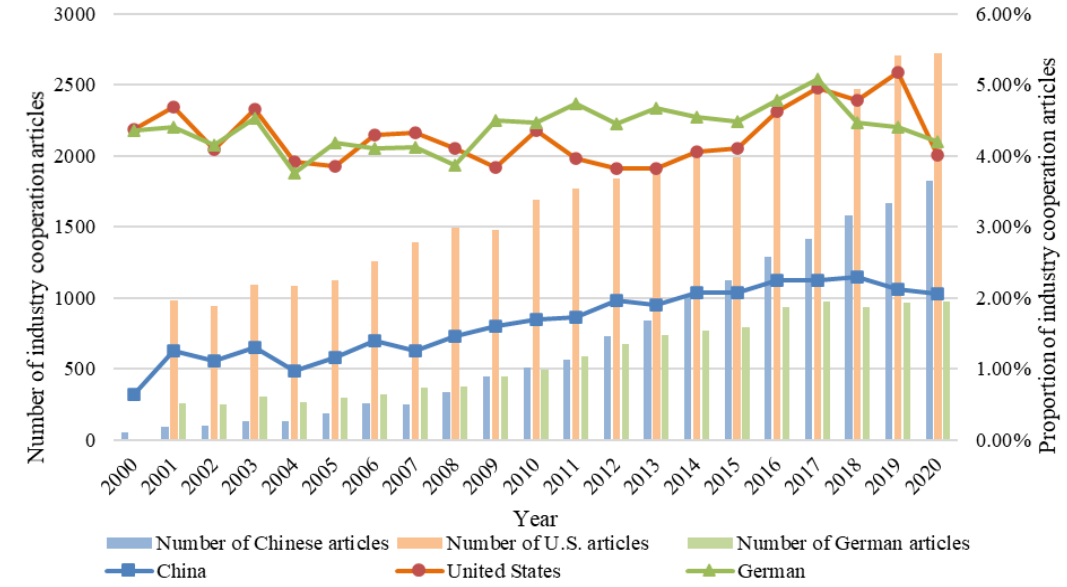


Figure 3. Temporal variation diagram of the quantity and proportion of industry co-authored articles of universities in China, Germany, and the United States from 2000 to 2020.

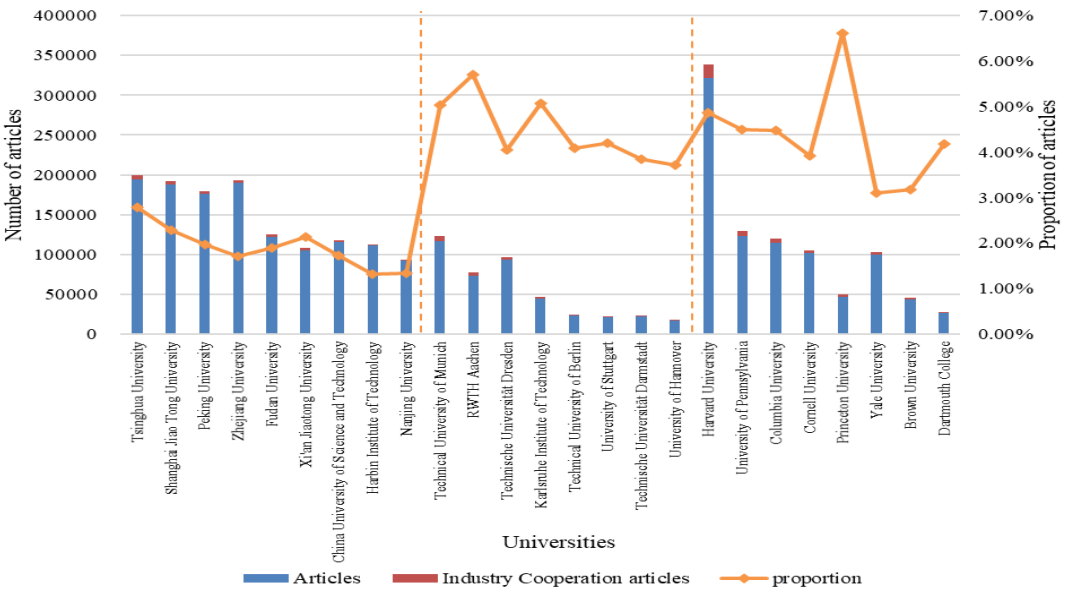


Figure 4. Diagram of the Quantity and Proportion of Industry Cooperation Articles of Chinese, German, and American Universities.

Figure 3 illustrates that the number of industryly co-authored articles in China, Germany, and the United States has generally increased each year. Notably, the

growth rate of industryly co-authored articles among Chinese universities has significantly accelerated since 2007, with the overall growth rate being the highest among the three countries.

A closer look reveals that the Chinese government implemented several policies to promote scientific and technological innovation around 2007. In 2006, China released the "Outline of the National Medium- and Long-Term Science and Technology Development Plan (2006-2020)")(P. s. R. o. China, 2006), which set forth objectives for advancing scientific and technological innovation. Subsequently, in 2010, the "Outline of the National Medium- and Long-Term Education Reform and Development Plan (2010-2020)" (P. s. R. o. China, 2010)explicitly stated the goals of enhancing higher education and strengthening scientific and technological innovation. This plan urged universities to enhance cooperation with all sectors of society and promote the transformation and application of research achievements.

It is evident that the combination of policy support and a conducive academic environment has fostered a collaborative relationship between Chinese universities and enterprises. In contrast, Germany has seen a stable trend in the number of industryly co-authored articles over the past five years. The number of co-authored articles between American universities and enterprises has fluctuated occasionally but generally exhibits an upward trend. This indicates that the cooperation between universities and enterprises in scientific and technological innovation in all three countries has become increasingly dynamic over the past two decades.

It can be observed from Figure 4 that while the number of industryly co-authored articles from Chinese universities has increased rapidly, its overall proportion remains relatively low compared to Germany and the United States. Over a span of 21 years, the average proportion of industryly co-authored articles from Chinese universities stands at only 1.91%, whereas both Germany and the United States exceed 4%. This suggests that, in terms of article co-authorship output, universities in Germany and the U.S. demonstrate stronger collaboration with enterprises.

Specifically, when examining the impact of the talent cultivation models of the United States and Germany on joint academic research between universities and enterprises, a notable example from the U.S. is the "I/UCRC" Industry-University Cooperative Research Center model (X. Wu, 2012). This model is supported by the National Science Foundation (NSF) and facilitates funding for general and fundamental research projects relevant to industry, thus encouraging collaborative research between industry and academia. In Germany, the prominent University Science Park model (Chen, Chu, & Hou, 2018) has been adopted. This approach creates an integrated cooperation system that links scientific research, education, and the economy, fostering active collaboration between scientific talent from universities and technical talent from enterprises, ultimately leading to the generation of numerous practical research outcomes.

Quantity and Temporal Variation in Cooperative Patents

Between 2000 and 2020, the total number of patents authorized by the C9 League universities in China exceeded 110,000. Tsinghua University alone had over 5,000 patents resulting from university-enterprise cooperation, significantly surpassing the numbers from Germany and the United States. However, the proportion of these collaborative patents was considerably lower than in Germany, accounting for less than one-third of the total. Additionally, the efficiency of patent conversion was relatively low. Table 2 provides an overview of the patent data for the sampled universities.

Table 2 Overall Situation of Patent Data of Sample Universities.

Country	University	University-Enterprise Cooperation Patents count	University-Enterprise Cooperation Patents Proportion
China	Tsinghua University	5960	22.79%
	Zhejiang University	1966	6.93%
	Peking University	1925	16.31%
	Shanghai Jiao Tong University	1510	8.41%
	Xi'an Jiaotong University	1094	7.65%
	Harbin Institute of Technology	535	3.04%
	Fudan University	321	5.68%
	Nanjing University	313	5.27%
	University of Science and Technology of China	219	4.99%
Germany	Dresden University of Technology	347	36.11%
	Technical University of Berlin	190	57.58%
	Technical University of Munich	158	43.89%
	University of Stuttgart	76	26.30%
	Darmstadt University of Technology	55	36.42%
	Brunswick Technical University	29	29.59%
	Karlsruhe Institute of Technology	9	23.08%
	RWTH Aachen University	4	33.33%
	University of Hanover	2	9.09%
	Harvard University	362	21.00%

United States	University of Pennsylvania	205	14.42%
	Yale University	173	23.60%
	Cornell University	169	17.16%
	Princeton University	107	14.60%
	Columbia University	106	7.28%
	Dartmouth College	35	9.54%
	Brown University	22	11.17%

Annual variation diagrams of the number and proportion of university-enterprise cooperation patents of different countries and universities were drawn with time and university as dimensions, as shown in Figures 5 and 6.

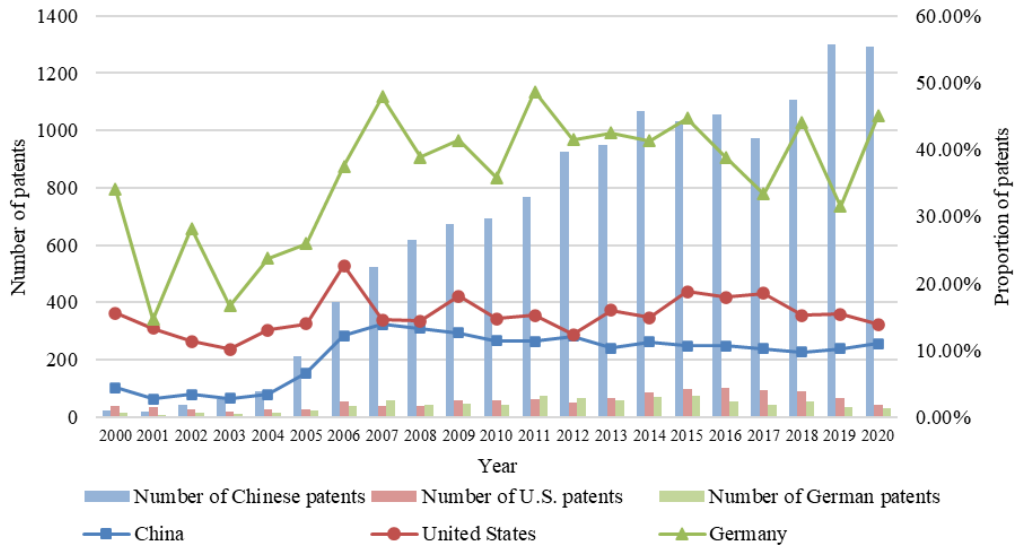


Figure 5. Annual Variation in the Quantity and Proportion of University-Enterprise Cooperation Patents among China, Germany, and the United States.

The figure shows that in Germany, the proportion of patents resulting from university-enterprise cooperation has been fluctuating at a relatively high level for the past 20 years. Since 2006, this proportion has consistently exceeded 30%. In contrast, the United States has maintained a more stable percentage, fluctuating between 10% and 25%. In China, the proportion of patents from university-enterprise cooperation increased steadily from 2004 to 2007. This rise can be attributed to the "Notice on the Establishment of National Technology Transfer Centers," issued in 2003 by the former State Economic and Trade Commission, the

Ministry of Education, and the Chinese Academy of Sciences (M. o. E. o. t. P. s. R. o. China, 2003). This initiative led to the establishment of multiple national-level technology transfer institutions, resulting in a brief surge in patent growth after 2004. However, the proportion of patents developed jointly by Chinese universities and enterprises remains relatively low. There is a pressing need for relevant policy guidance, incentive measures, and a robust protection mechanism to address this issue.

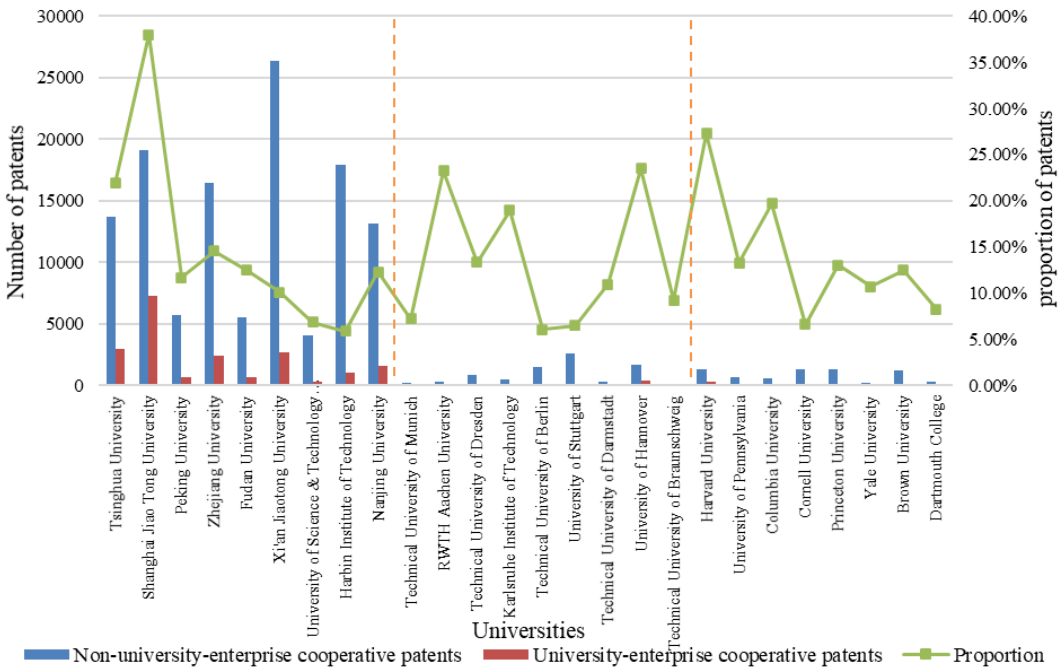


Figure 6. Quantity and Proportion of University-Enterprise Cooperation Patents of Chinese, German, and American Universities.

Certain scholars have delved into the factors contributing to the relatively low conversion efficiency of patents held by Chinese universities, and the main aspects are as follows. To begin with, the patents of Chinese universities generally exhibit deficiencies in both quality and practicality (JuJie Guo, He, & Huang, 2007; D. Liu, 2018). Chinese institutions of higher learning have indeed filed a substantial number of patents based on their scientific research endeavors. Nevertheless, these research projects frequently commence from academic topics and tend to overlook the actual market trends and the specific requirements of enterprises. As a consequence, the resultant patents face significant hurdles in terms of marketability. Most of these patents have not been subjected to production experiments and remain confined to

the laboratory stage, rendering it arduous for enterprises to integrate them into their actual business operations.

Secondly, Chinese universities notably lack professional patent management and conversion institutions (Zhang & Huang, 2011). The initiation of patent conversion activities in Chinese universities has been relatively tardy. The vast majority of university research management departments are primarily engaged in the routine tasks of patent application and daily patent management. These departments are bereft of the necessary capabilities for conducting application evaluations of the patent market, which impedes their ability to effectively facilitate the conversion of patent achievements. Concurrently, both Chinese universities and enterprises are found to be deficient in corresponding patent conversion incentive mechanisms(Nonaka, 1994). The majority of universities have not incorporated patent conversion into their strategic agendas. Moreover, the process of achievement conversion demands a substantial investment of energy and financial resources. University faculty members lack the requisite motivation, and enterprises are disinclined to assume risks and allocate significant amounts of capital.

In contrast, Germany and the United States have established increasingly sophisticated achievement conversion systems. In the United States, most universities are equipped with technology transfer offices, and there are specialized agencies dedicated to conducting commercial research and identifying suitable partners. In Germany, the technology transfer funds of research universities have garnered robust support from the government, enterprises, and public welfare organizations. The government has also established multiple science and technology centers to offer free consulting services to enterprises, thereby effectively promoting the conversion of scientific research achievements(Sun, Liu, & Xu, 2016).

The study further integrates science and technology indicators to comprehensively analyze university-industry collaboration. A comprehensive analysis of the data of co-authored articles and cooperation patents between universities and enterprises shows that there are certain differences among China, Germany, and the United States in the proportion of industry co-authored articles and the proportion of university-enterprise cooperation patents. German universities perform better in both indicators and are more active in innovation cooperation with enterprises. Figure 7 is a scatter plot of the data of co-authored articles and patents of different universities and enterprises. Different shapes and colors in the figure represent different countries, and two line segments are used to mark the mean values of the relevant proportions.

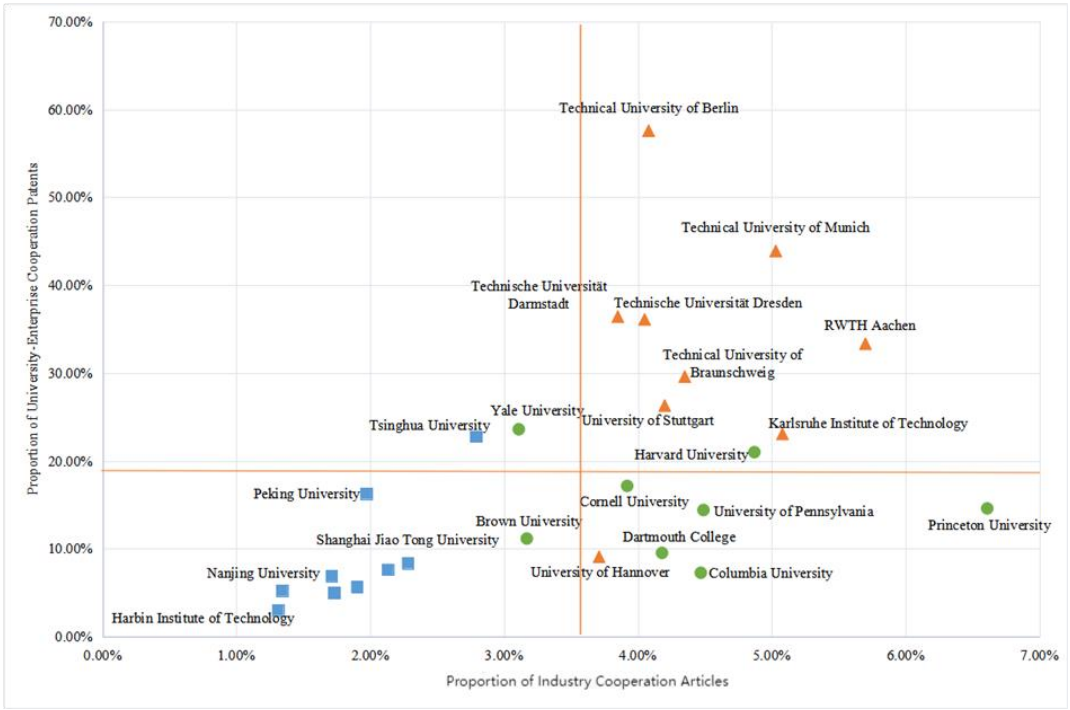


Figure 7. Scatter Plot of the Proportion of Industry Co-authored Articles and the Proportion of University-Enterprise Cooperation Patents.

Note: The squares represent Chinese universities, the triangles represent German universities and the circles represent American universities.

The proportion of industryly co-authored articles and university-enterprise cooperation patents at the Technical University of Munich and the Technical University of Berlin in Germany is significantly higher than that of other universities. This indicates that these two institutions have clear advantages in innovation collaboration with enterprises. In the United States, universities perform better in terms of industryly co-authored articles, with most institutions having a proportion that exceeds the average. However, their performance concerning university-enterprise cooperation patents is relatively mediocre, with only Yale University and Harvard University surpassing the average level.

In contrast, the level of innovation cooperation between Chinese universities and enterprises is lower compared to their counterparts in Germany and the United States. Only Tsinghua University displays a proportion of university-enterprise cooperation patents that exceeds the average, highlighting a stark contrast to the high number of authorized patent data from Chinese universities. This suggests that while Chinese universities possess strong capabilities in innovative research, many of their

innovative achievements and authorized patents remain unutilized and have not fully transitioned into practical applications.

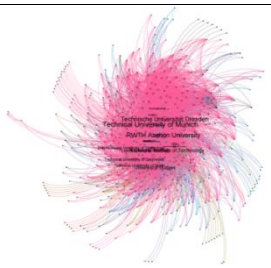
Despite their significant experience and accomplishments in scientific research, Chinese universities encounter substantial challenges in cooperation in scientific and technological innovation and in converting these achievements into applicable solutions. Thus, finding ways to enhance university-industry cooperation and improve the efficiency of converting innovative achievements has become an urgent issue that needs immediate attention.

This study analyzes the network structure characteristics of university-enterprise cooperation in China, Germany, and the United States from the perspective of cooperation networks, exploring the performance of indicators such as the scale, intensity, and average degree of university-enterprise cooperation across these different countries.

Network Analysis based on Co-authored articles

Upon examining the cooperation network diagram, it is evident that the university-enterprise collaboration in the field of published articles across the three countries generally exhibits a galaxy-like network structure. In this network, research institutions, large companies, and high-tech enterprises often serve as the core nodes alongside universities, with most nodes gathering around universities as central hubs. This indicates that the collaborative relationships among universities are generally closer than those between universities and enterprises. Additionally, the cooperation network diagrams for Germany and the United States show a greater diversity of nodes. Notably, the number of enterprises co-authoring articles with German universities is the highest, while there is a comparatively smaller number of enterprises collaborating with Chinese universities. The following table presents the cooperation network diagrams and relevant structural data pertaining to co-authored articles from universities and enterprises in China, Germany, and the United States.

Table 3. Cooperation Network Diagrams of Co-authored articles of Universities and Enterprises in China, Germany, and the United States.

	China	Germany	United States
Cooperation Network Diagram			
Network Node Scale	155	523	208
Network Edge Number	1476	4545	1950
Average Degree	19.05	17.38	18.75
Average Clustering Coefficient	0.662	0.793	0.797

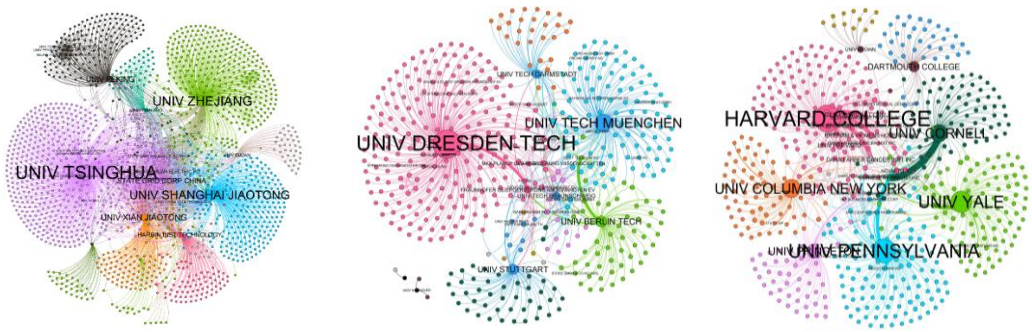
Note: Only nodes with a frequency greater than 10 are shown in the figure for the convenience of presentation.

In terms of the overall scale of cooperation networks, the collaboration between German universities and enterprises is the largest. Both the number of partnering enterprises and the frequency of cooperation are higher than in the other two countries. This trend is closely linked to Germany's long-standing emphasis on university-industry collaboration. The German government has implemented various policies to support and enhance this cooperation. For instance, the "Employee Invention Law" stipulates that 30% of the income generated from the patent conversion of employee inventions will be awarded to the inventors. Additionally, in 2014, the German government launched the "High-Tech Strategy 2025," which identifies university-industry cooperation as a key component aimed at improving Germany's innovation capacity and scientific and technological competitiveness (STIPCOMPASS, 2018).

The average degree index indicates the overall connection status of all nodes in the network diagram, while the average clustering coefficient measures the degree of clustering among these nodes. The clustering coefficients for Germany and the United States are higher than that of China, and their average degree is slightly lower. This suggests that the collaboration between these two countries and their enterprises in terms of article co-authorship is more cohesive, and the partnerships between universities and enterprises are more balanced. In contrast, the cooperation network of Chinese universities shows a relatively high average degree but a low average clustering coefficient. This indicates that the collaboration among Chinese universities and research institutions in co-authorship is not well balanced. Analyzing specific co-authorship data reveals that some Chinese universities tend to cluster with other universities or research institutions. Prominent universities and research institutions hold significant positions in article co-authorship, leading to concentrated collaboration among them. Meanwhile, Chinese enterprises have a comparatively minor role in scientific research, with fewer connections to the core universities in the network diagram. This results in a cooperation pattern that predominantly features an aggregation of resources among universities and research institutes.

Network Analysis Based on Cooperative Patents

There are similarities in patent cooperation among sample universities in China, Germany, and the United States. Each university has a fixed group of cooperative enterprises, and these enterprises have established close cooperation relationships with specific universities to jointly promote scientific and technological innovation research. Figure 8 shows the cooperation network diagrams of patents of universities and enterprises in China, Germany, and the United States.



China Germany United States

Figure 8. Cooperation Network Diagrams of Patents of Universities and Enterprises in China, Germany, and the United States.

Note: Only nodes with a cooperation frequency greater than 1 are shown in the figure for the convenience of presentation.

The figure illustrates that the collaboration between universities and enterprises tends to cluster around individual universities. Each university has a specific group of partner enterprises with whom they have formed close cooperative relationships to advance scientific and technological innovation research. However, this collaboration often appears somewhat limited; most enterprises establish a partnership with only one university and do not reach out to others afterward.

One possible explanation for this is that universities, as knowledge producers, offer unique and diverse resources that many enterprises cannot replicate (Fukugawa, 2013). This creates a situation where multiple enterprises compete for collaboration with universities, but due to the distinct research areas and technical expertise of each institution, enterprises ultimately select the university that best aligns with their needs and capabilities. This results in a one-to-many cooperation model between universities and enterprises.

Additionally, the enterprises that closely collaborate with leading universities—whether domestically or internationally—are typically well-established and relatively large organizations. This suggests that such enterprises prioritize partnerships with top-tier universities, viewing them as vital for their technological innovation and research and development efforts. Furthermore, it has been observed that universities also engage in patent cooperation and joint research. This collaborative pattern fosters the sharing of resources and knowledge among universities, further enhancing technological innovation.

Table 4 The Top Two Enterprises with the Highest Cooperation Frequency of Each University.

Country	University	Enterprise
China	Tsinghua University	Shenzhen Foxconn Precision Group; Yida Technology Co., Ltd.
	Zhejiang University	State Grid Zhejiang Electric Power Co., Ltd.; Zhejiang Nanhu Co., Ltd.
	Peking University	Peking University Founder Group Co., Ltd.; Beijing Chuangshitong Technology Co., Ltd.
	Shanghai Jiao Tong University	State Grid Corporation; Huawei Technologies Co., Ltd.
	Xi'an Jiaotong University	State Grid Corporation; Xi'an Ruite Rapid Manufacturing Engineering Co., Ltd.
	Harbin Institute of Technology	State Grid Corporation; Harbin Institute of Technology Ruichi Technology Co., Ltd.

	Fudan University	Shanghai iQIYI Innovation Center Co., Ltd.; Huawei Technologies Co., Ltd.
	Nanjing University	Jiangsu Enju Environmental Protection Technology Co., Ltd.; Suzhou Nanzi Sensing Technology Co., Ltd.
	University of Science & Technology of China	Huawei Technologies Co., Ltd.; State Grid Corporation
Germany	Technical University of Dresden	Fraunhofer Institute for Applied Technology Promotion; Novald Company
	Technical University of Berlin	Deutsche Telekom AG; Fraunhofer Institute for Applied Technology Promotion
	Technical University of Munich	Bavarian Motor Works; Lanxess AG
	University of Stuttgart	Audi AG; Garnier Construction Machinery Company
	Technical University of Darmstadt	Fraunhofer Institute for Applied Technology Promotion; Deutsche Telekom AG
	Technical University of Braunschweig	Fraunhofer Institute for Applied Technology Promotion; Innovation Laboratory
	Karlsruhe Institute of Technology	Karlsruhe Research Center GmbH; Fraunhofer Institute for Applied Technology Promotion
	RWTH Aachen University	ASML Netherlands; ASML Company
	University of Hannover	BIOTRONIK SE & Co. KG; Braun Company
United States	Harvard University	Broad Institute; Dana-Farber Cancer Institute, Inc.
	University of Pennsylvania	Novartis Technologies Ltd.; INOVIO Biopharmaceuticals, Inc.
	Yale University	Yale University Corporation; Regeneron Pharmaceuticals, Inc.
	Cornell University	Cornell Research Foundation, Inc.; Nestlé Science and Technology Co., Ltd.
	Princeton University	Universal Display Corporation; Momentive Performance Materials, Inc.
	Columbia University	AT&T Inc.; Sony Corporation; Dana-Farber Cancer Institute, Inc.
	Dartmouth College	Maskoma Corporation; Immunex Corporation
	Brown University	Xerox Network Services; League for Sustainable Energy, LLC

By analyzing the top two enterprises with the highest frequency of cooperation from each university (as shown in Table 4), it is evident that there are distinct characteristics in patent collaboration between universities and enterprises across different countries. The cooperation network involving Chinese universities and enterprises is notably richer. Led by Tsinghua University, Zhejiang University, and Shanghai Jiao Tong University, each institution has established its own unique network of partnerships. A closer examination of the enterprises that collaborate most frequently with these Chinese universities indicates that each university tends to partner with companies located in the same region or those with which the university shares (Ding et al., 2010). For instance, the enterprises with the closest ties with Tsinghua University, Zhejiang University, and Peking University are the Shenzhen Foxconn Precision Group, State Grid Zhejiang Electric Power Co., Ltd., and Peking University Founder Group Co., Ltd.

Each German university has formed its own unique cooperation cluster group. An analysis of specific patent cooperation data shows that the institutions that collaborate most frequently with the German Universities of Technology League are primarily off-campus public research institutions, as well as well-known enterprises both within Germany and internationally. For example, the top three institutions with the highest frequency of patent cooperation with German universities are Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V., Deutsche Telekom AG, and AUDI AG. This trend aligns with the structure of the German innovation system and government innovation policies. Germany has developed a scientific and technological innovation system with universities, public research institutions, and enterprises serving as its three pillars. The Fraunhofer Society is one of the most representative research institutions in this system. The German innovation framework clearly defines the roles and operational mechanisms of each entity and promotes collaborative efforts among these three innovation sectors based on local conditions. Germany has established a stable cooperation platform that fully mobilizes the scientific and technological innovation capabilities of universities and enterprises. This has improved the efficiency of converting scientific and technological achievements into practical applications, providing a strong foundation for fostering national development.

Institutions that frequently collaborate with American universities are primarily research funding organizations established by universities and various research enterprises. Examples include the Cornell Research Foundation, Inc., the Broad Institute, and the Dana-Farber Cancer Institute, Inc. In managing university-industry partnerships, many American universities set up dedicated technology management offices to facilitate the transformation of scientific and technological achievements (Yang, 2011). Additionally, some universities create separate management entities to handle technology transfer and intellectual property matters, thereby promoting

scientific research and technological innovation. These institutions operate independently from the university's main administration. By providing commercial services, they support university operations while maintaining greater authority and autonomy, which can lead to more efficient transformation of scientific research achievements.

Discussion

By conducting a bibliometric analysis of university-industry cooperation data from 26 top universities in China, Germany, and the United States, and constructing a cooperation network, we have uncovered the innovation cooperation patterns and characteristics of universities and enterprises in these three countries. The key conclusions are as follows: First, universities in Germany and the United States demonstrate better performance in terms of innovation achievements in collaboration with enterprises. While Chinese universities hold the largest number of authorized patents, the proportion of patents resulting from university-industry cooperation is relatively low, and the conversion rate of these innovation achievements is also not high. Second, universities and enterprises typically form a galaxy network structure when examining the cooperation network among the three countries. In addition to university nodes, research institutions, large enterprises, and high-tech enterprises often act as core nodes in the network, with all nodes gravitating toward the universities at the center.

These findings indicate that university-industry innovation cooperation in China has achieved notable success over the past two decades. However, compared to the cooperative frameworks in Germany and the United States, China still faces significant challenges in enhancing university-industry collaboration and the conversion of scientific and technological achievements, indicating ample room for improvement. Specifically, universities in Germany and the United States not only possess mature cooperation models and operational systems with enterprises but have also developed a relatively comprehensive ecosystem regarding achievement conversion mechanisms, policy support, and enterprise involvement. Although China has made some progress, it needs to exert further effort to deepen university-industry cooperation and improve the efficiency and quality of converting scientific and technological achievements.

In the case of the United States, while American universities excel in industry cooperation in research articles, they still exhibit weaknesses in university-industry cooperation patents. Moreover, compared to universities, the enterprises they collaborate with are relatively limited, with many enterprises maintaining stable and unchanging partnerships with a select few universities.

This study offers a comprehensive analysis of the university-industry collaboration status among C9, TU9, and Ivy League universities from a science and technology

perspective. Future research could delve deeper into the impact of national science and technology innovation policies. The policy directions, government support priorities, and strategic frameworks of different countries significantly shape the models of university-industry collaboration.

Additionally, future studies might employ more advanced research methods, such as deep learning and text mining, to identify hidden patterns and relationships. Incorporating dynamic network analysis could also be beneficial, as it would investigate how the collaboration networks between universities and industries evolve over time, providing a fresh perspective for assessing the long-term outcomes of these partnerships.

Acknowledgment

The authors would like to acknowledge support from the National Natural Science Foundation of China (Grant Nos. 72374162, L2324105, and L2424104) and the National Laboratory Centre for Library and Information Science at Wuhan University.

References

Chen, H., Chu, G., & Hou, J. (2018). A Comparative Analysis on the Domestic and Foreign Modes of Cultivating Innovative Talents in Industry-University-Research Institute Cooperation. *Forum on Science and Technology in China*(01), 164-172.

China, M. o. E. o. t. P. s. R. o. (2003). Notice on the Establishment of National Technology Transfer Centers by the State Economic and Trade Commission, Ministry of Education, and Chinese Academy of Sciences. Retrieved December 30, 2024 from http://www.moe.gov.cn/jyb_xxgk/gk_gbgg/moe_0/moe_9/moe_35/tnull_427.html.

China, P. s. R. o. (2006). National Medium and Long-term Program for Science and Technology Development (2006-2020). Retrieved December 30, 2024, from http://www.gov.cn/gongbao/content/2006/content_240244.htm.

China, P. s. R. o. (2010). National Medium and Long-term Plan for Education Reform and Development (2010-2020) Retrieved December 26, 2024, from http://www.moe.gov.cn/srcsite/A01/s7048/201007/t20100729_171904.html.

China, P. s. R. o. (2021). Outline of the 14th Five-Year Plan for National Economic and Social Development and the Long-Range Objectives Through the Year 2035. Retrieved December 30, 2024, from http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm

China, P. s. R. o. (2022). Hold high the great banner of socialism with Chinese characteristics and strive in unity to build a modern socialist country in an all-round way - Report at the 20th National Congress of the Communist Party of China. Retrieved January 3, 2025

Dang, Q. T., Jasovska, P., Rammal, H. G., & Schlenker, K. (2019). Formal-informal channels of university-industry knowledge transfer: the case of Australian business schools. [Article]. *Knowledge Management Research & Practice, 17*(4), 384-395.

Ding, L., Huang, L., & Guo, K. (2010). Research on the Industry-Academia Cooperation Model in Vocational Colleges and Institutes. *West China Development*(4), 112-112.

Etzkowitz, H., & Leydesdorff, L. (1996). A triple helix of academic-industry-government relations: Development models beyond 'capitalism versus socialism'. [News Item]. *Current Science, 70*(8), 690-693.

Foundation, N. S. (2018). U.S. and International Research and Development: Founds and Alliances. Retrieved December 27, 2024, from http://www.nsf.gov/statistics/seind02/c4/c4s3.htm.

Fukugawa, N. (2013). University spillovers into small technology-based firms: channel, mechanism, and geography. [Article]. *Journal of Technology Transfer, 38*(4), 415-431.

Guo, J., He, M., & Huang, Y. (2007). The Analsis on the Transfomation of the Patent Technology in Universities & Colleges and It's Commercialized Countemm easures. *Technology and Innovation Management*(03), 80-83.

Guo, J., Xie, F., Wang, H., & Wang, M. (2019). Research on the Dynamic Evolution of Cooperative Networkbased on Industry-University Joint patent Data——An Analysis of" Double First-Class" Universities. *Science & Technology Progress and Policy, 36*(17), 1-10.

Hu, J., Zhu, G., & Ma, Y. (2011). System Dynamic Analysis on Influencing Factors of Industry-University-Research Cooperation in Open Innovation Context. *Science of Science and Management of S.& T, 32*(08), 49-57.

Kenney, M., & Patton, D. (2009). Reconsidering the Bayh-Dole Act and the current university invention ownership model. *Research Policy, 38*(9), 1407-1422.

Kwon, K. S., Park, H. W., So, M., & Leydesdorff, L. (2012). Has globalization strengthened South Korea's national research system? National and international dynamics of the Triple Helix of scientific co-authorship relationships in South Korea. [Article]. *Scientometrics, 90*(1), 163-176.

Liu, D. (2018). Transformation of University Patents in China： Problems and Strategies. *Tianjin Science & Technology, 45*(06), 1-5.

Liu, F., Ma, R., & Jiang, N. (2011). Research on Evolutionary Paths of Industry — University — Research Institute Networks of Patent Collaboration Based on the“985 Universities”. *China Soft Science*(07), 178-192.

Nonaka, I. (1994). A DYNAMIC THEORY OF ORGANIZATIONAL KNOWLEDGE CREATION. [Article]. *Organization Science, 5*(1), 14-37.

Rast, S., Khabiri, N., & Senin, A. A. (2012, Jan 13-15). *Evaluation Framework for Assessing University-Industry Collaborative Research and Technological Initiative.* Paper presented at the International Conference of the Asia-Pacific-Business-Innovation-and-Technology-Management-Society, Pattaya, THAILAND.

STIPCOMPASS. (2018). High-Tech Strategy 2025. Retrieved January 2, 2024, from https://stip.oecd.org/moip/case-studies/1.

Sun, Z., Liu, C., & Xu, R. (2016). Research University Technology Transfer Optimization Design Under The Perspective of System Comparison. *Science and Technology Management Research, 36*(20), 72-77.

Wang, S. (2020). Impact of university-industry collaboration on overall technology innovation performance of the university. *Journal of Innovation and Entrepreneurship Education, 11*(02), 1-6.

Wang, X., Wang, H., & Liu, L. (2005). Research on Industry-Academia-Research Alliance Models and Selection Strategies. *China University Science & Technology*(11), 64-67.

Wu, R., Liu, S., & Li, Z. (2021). Research on Knowledge Transfer Mechanisms in Industry-Academia-Research Collaborative Innovation. *Technology and Industry Across the Straits, 34*(06), 13-18.

Wu, X. (2012). Construction of Collaborative Innovation Alliance Involving Production，Teaching and Research and Experiences: I/UCRC Model. *China Higher Education Research*(04), 47-50.

Xiao, Y. (2016). Lessons and References for China from Typical Cases of Industry-Academia-Research Cooperation in Developed Countries: The Example of Germany's Dual System. *China University Science & Technology*(10), 43-45.

Yang, G. (2011). A Survey of the Transfer and Transformation of S&T Achievements in the United States. *Science & Technology for Development*(09), 87-93.

Zhang, P., & Huang, X. (2011). The Current Status, Issues, and Development of Patent Technology Transfer in Chinese Universities. *China Higher Education Research*(12), 34-37.

https://doi.org/10.51408/issi2025_001

Text-based Classification of All Social Sciences and Humanities Publications Indexed in the Flemish VABB Database

Cristina Arhiliuc[1], Raf Guns[2], Tim C. E. Engels[3]

[1] *cristina.arhiliuc@uantwerpen.be,* [2] *raf.guns@uantwerpen.be,* [3] *tim.engels@uantwerpen.be*
Centre for Research and Development Monitoring (ECOOM), University of Antwerp,
Middelheimlaan 2, 2020 Antwerp (Belgium)

Abstract

This research describes and evaluates a new methodology for classifying peer-reviewed publications based on the textual metadata available. The methodology is developed for application to the Flemish database for Social Sciences and Humanities (VABB-SHW) and could also be applied in similar databases. To build the classification model, we fine-tune the SSCI-SciBERT model with textual features of journal articles (journal titles, publication titles and abstracts) from Web of Science corresponding to the time period 2000-2022 that is covered by VABB-SHW. We experiment with different feature combinations to replicate the lack of abstracts or the publication channel for a proportion of publications in the target dataset. We conclude that the combined model, trained to handle various combinations of textual features, achieves similar results to feature(s)-specific models, while being more convenient to use. Then, to be able to apply the fine-tuned SSCI-SciBERT to the multilingual VABB-SHW dataset, we translate its data to English using gpt-4o-mini. As the VABB-SHW data is mostly unlabelled at the publication level and covers more publication types than the training dataset, we conduct a separate evaluation for the quality of the classification at the publication type level both by using the prior existing classification (for books and book chapters with generic names) and by comparing it with a manually classified sample of the data and evaluating the quality of the model classification. The model achieves a F1-score of 55% on the VABB-SHW test dataset, with publication type an impacting factor.

Introduction

The goal of this research is to propose a new method for the paper-level, text-based multilabel classification of research publications. The proposed approach is applied to the VABB-SHW database, which stores publications (co-)authored by researchers from Social Sciences and Humanities departments at Flemish universities. The models developed through this method can also be used for the classification of other scholarly and scientific texts. Moreover, this paper specifically examines how well data from journal articles transfers to other types of publications, namely conference proceedings, books, and book chapters.

National bibliographic databases have been created in several countries and regions to offer a comprehensive resource for studying and monitoring the research publications produced in a country or region (Sīle et al., 2018). Among other fields, such databases are especially relevant in the Social Sciences and Humanities. These fields are in their nature and research tradition more locally anchored and typically less well-covered in international citation indexes (Archambault et al., 2006; Sīle et al., 2017, 2018; Sivertsen, 2016; Sivertsen & Larsen, 2012). Although national bibliographic databases are usually more comprehensive, due to their local coverage

they often lack citation information, which precludes classifying individual papers according to discipline making use of their positioning in the citation network as it is often done for paper-level classification using Web of Science (Perianes-Rodriguez & Ruiz-Castillo, 2017; Waltman & van Eck, 2012). Hence here we rely on natural language processing of the textual metadata of the publications to classify them to disciplines.

The VABB-SHW database has been implemented in 2008 to complement the Web of Science (WoS) data, which has a low coverage in the Social Sciences and Humanities with the purpose of implementing a fairer performance-based research funding system (Verleysen et al., 2014). The original classification in the database is an organizational classification, i.e. a classification that labels each paper with the discipline(s) of the unit(s) of its (Flemish) authors. This classification gives information about *who* writes the papers from the database. Later, a new classification has supplemented the organizational one: the cognitive channel-based classification that assigns to each publication the discipline(s) of the journal, conference proceeding, book or book series that it originates from (Guns et al., 2018). This cognitive classification provides information regarding where the publications written by Flemish SSH researchers are published. Finally, the paper-level classification presented in this paper supplements the existing two classifications and provides a more fine-grained classification of all the publications included in the VABB-SHW. It answers the question "what *disciplines* do the SSH researchers in Flanders contribute to?".

To train a model for the classification task, we require labelled data, which is not available at the publication level in the VABB-SHW database. Therefore, we use WoS data to train and evaluate different model configurations before applying them to our local database, relying on the classification of references of a paper in WoS to infer the final ground truth. While several studies have identified issues with the accuracy and consistency of WoS classifications (Aviv-Reuven & Rosenfeld, 2023; Milojević, 2020; Singh et al., 2020; Wang & Waltman, 2016) they also acknowledge that WoS remains one of the more reliable options for large-scale classification tasks. At an aggregate level, we consider it to provide a sufficiently robust foundation for this research. The WoS Science categories are mapped to an extended version of the OECD FoRD classification scheme (OECD, 2015) as this scheme is used for all classifications in the system.

This paper builds on our previous work in which we explored appropriate models for text-based classification of publications (Arhiliuc et al., 2024). On the basis of those previous findings, we select the SSCI-SciBERT model (Shen et al., 2022).

Throughout this paper we answer the following questions:

1. Which ground truth labelling strategy represents the data the best, while keeping the distribution of the number of labels to what we are currently expecting in our database?
2. Which strategy, accounting for the varying availability of distinct textual features, yields the best classification results?
3. How well does the knowledge extracted through model fine-tuning from WoS journal articles transfer to non-WoS articles and to other publication types?

In the following parts we first introduce the data, both the WoS data used for model fine-tuning for the classification task and the final application – the VABB-SHW data. Secondly, we explain the methodology and the evaluation procedure for the models. Thirdly, we present the results. We end with conclusion and discussion of the overall implications of this research and further work to be done.

Data description

This project uses two datasets. First, due to the unlabelled nature of the VABB-SHW data, Web of Science (WoS) data has been used to fine-tune the models for the task of classification of the scientific literature. Then, the pretrained models evaluated on the WoS data are applied on the local VABB-SHW database. This section describes the characteristics of both datasets.

WoS data

Web of Science is an international database that indexes peer-reviewed publications and provides extensive metadata. This includes publication titles, years, channels (e.g., journals, conference proceedings, books, or book series), disciplines (referred to as "science categories") assigned at the channel level, and citation information. In this study, we use data from three WoS indices - the Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and Arts & Humanities Citation Index (AHCI).

We have previously run classification experiments on WoS data for the year 2022 (Arhiliuc et al., 2024). However, the disciplines from Social Sciences and Humanities were often underrepresented, which might have caused worse results overall for them. Due to the nature of the current classification task, where we aim to classify all publications written by researchers from Social Sciences and Humanities included in the VABB-SHW, it is essential to have a good coverage of those fields. We have therefore extended the dataset to include publications from multiple years in the range of years 2000-2022 represented in VABB-SHW. More precisely, we have extracted all the journal articles indexed in the Web of Science from the years 2002, 2006, 2010, 2014, 2018, and 2022. This dataset contains 7 973 222 publications.

The subject categories from WoS are then mapped to OECD FoRD categories (OECD, 2015) with an extension at the level of humanities: "History and archaeology" is split into "History" and "Archaeology", "Languages and literature" into "Languages and linguistics" and "Literature", and "Philosophy, ethics and religion" into "Philosophy and ethics" and "Religion" as this is the classification used in VABB-SHW. Three other disciplines are however excluded due to not being present in the mapping scheme: "Other natural sciences", "Other medical sciences", "Agricultural biotechnology". Additionally, multiple science categories marked as multidisciplinary in the WoS classification are mapped to "Multidisciplinary" discipline that is however less relevant when classifying at the publication level than at the channel level, so it will therefore not be used in this study.

VABB-SHW data

For this study, we are using the 14th edition of the VABB-SHW database that contains publications written by scholars from SSH departments from Flemish universities in the years 2000-2022, both peer-reviewed and not peer-reviewed. The metadata used for this research includes the publication title, abstract, and channel title. The channel depends on the publication type: journals for journal articles, conferences for conference proceedings papers, books for book chapters, and book series for authored or edited books.

Table 1 presents the specificities for each of the five publication types available in VABB-SHW. However, for the purpose of evaluation, they have been grouped into three groups based on their characteristics:

1. Journal articles are conference proceedings – are characterized by a higher availability of the abstracts in the database, more general channel titles and specific publication title
2. Books as author and books as editor ultimately represent the same entity type: books and have been grouped as such
3. Book chapters – can be both specific and general and normally make sense mostly in combination with the channel title.

Table 1. Availability of textual features and publication language for different publications types in VABB-SHW 14.

Type	*Count*	*Abstract*	*Channel title*	*Publication title*	*English*	*Dutch*
Journal articles	170 418	70 869 (41.59%)	170 340 (99.95%)	170 418 (100%)	**61.39%**	33.59%
Authored books	16 295	3 318 (20.36%)	5 838 (35.8%)	16 295 (100%)	25.04%	**64.04%**
Edited books	11 843	1 824 (15.40%)	6 297 (53.17%)	11 843 (100%)	45.88%	42.77%
Book chapters	74 071	9 212 (12.44%)	74 043 (99.96%)	74 071 (100%)	45.15%	41.77%
Conference proceedings papers	12 851	6 187 (48.14%)	12 849 (99.98%)	12 851 (100%)	**83.10%**	9.00%

Table 1 highlights several key characteristics of the available textual features in the dataset. The publication title is fully available across all publication types. The channel title (i.e., journal, conference proceedings, book, book series) is available, with some exceptions, for journal articles, conference proceedings papers and book chapters, but is less commonly available for books (both authored and edited). Abstracts, as previously mentioned, are primarily associated with journal articles and conference proceedings. Moreover, conference proceedings papers and journal articles are mostly in English, while books as author are mostly in Dutch and edited

books and book chapters have similar numbers for English and Dutch with a smal share of publications in other languages. These differences may lead to variations in the quality of classification.

Methodology

The current research has two main parts.
The first part uses the labelled WoS data to search for the right model structure and configuration to fit our classification requirements. The requirements are based on similar previous tasks and the characteristics of the VABB-SHW data: multilabel classification, in preponderantly one to three disciplines, able to provide optimal results based on the availability of the textual data representing a publication.
The second part focuses mainly on the VABB-SHW and covers the preparation of the VABB-SHW data for the classification, the application of the strategy designed in the first part and the evaluation of the classification.

Part 1: Model selection

Thresholds

Determining relevant ground-truth labels for the WoS data is fundamental for this research. The ground-truth classification for a specific publication is deduced from the distribution of disciplines in the reference list. However, this raises the question: what proportion of the references of a paper should be in a specific discipline to assume that the discipline is representative of the content of the paper?
In the ECOOM-Biblio-Antwerp team that is responsible for the maintenance and analysis of the VABB-SHW database, we have an annual task of manual classification at a journal, conference and book level. This is done to enrich the existing channel-based cognitive classification when no data regarding those channels has been automatically found in external sources. One of the guidelines for that task is limiting the number of disciplines to a maximum of three. Based on that, in a previous study of classification methods for journal articles (Arhiliuc et al., 2024), we have selected the threshold of 0.3 as most publications get classified in up to three disciplines with relatively few publications being classified in no discipline or more than five disciplines. In this study however, we aim on a more methodical analysis of the appropriate threshold that is going to happen in two steps:

1. Analysis of the distribution of the number of disciplines assigned to publications using thresholds varying from 0 to 1 with a 0.05 interval between them. The Multidisciplinary discipline to which multiple multidisciplinary science categories map, has been removed from this analysis, resulting in a few outliers having 0 disciplines even at threshold 0. The goal of this analysis is to select the thresholds that position most publications in 1 to 3 disciplines, which is what we are aiming for. More precisely, we are looking for the thresholds that have more than 90% of the publications in 1 to 3 disciplines to maximize the number of publications available for the creation of the train, validation and test datasets.
2. As a proxy of how representative the labels are of the data, train, validation and test datasets are created for each of the selected thresholds and then SSCI-

SciBERT is assigned with the task of classifying the publication into disciplines based on their abstracts. Small variations in the results among thresholds should not be viewed as significant as due to the variation in number of disciplines per publication for each threshold, the datasets are distinct among thresholds, which can have an impact on the result.
The optimal threshold is selected based on the distribution of number of disciplines and the F1-score on the test datasets in the second step.

Data partitioning

For all the experiments in this part, the train, validation and test datasets are selected to be as balanced as possible across disciplines given the multilabel nature of the classification. More specifically, we aim to select 500 examples per disciplines for the test dataset, 500 examples per discipline for the validation dataset and 10 000 examples for the train dataset if available.
To test various model configurations after the choice of the threshold (see *Thresholds*), we partitioned the data into separate train, validation, and test sets, ensuring no overlap of journals across the three sets to prevent leakage when using the journal names. Due to data availability challenges in certain Social Sciences and Humanities disciplines and the constraints of this partitioning approach, we prioritized maximizing the number of publications in the training set for underrepresented disciplines. To achieve this, we allocated publications from the least represented journals to the test and validation sets, avoiding the placement of journals with large numbers of examples in these smaller sets, where many examples would go unused. While this approach ensures an efficient use of available data for training, it reduces the randomness of partitioning.
Moreover, a second drawback of this method must be considered: by distributing journals among the three datasets, it is possible that no set fully captures the diversity of the disciplines, as distinct journals might focus on different aspects of the field.
Additionally, if the goal is to classify new publications, having a greater variety of journals in the training set could enhance classification quality, as the model benefits from learning discipline-specific patterns associated with that journal. Therefore, while datasets with no journal overlap across the three sets provide an opportunity to test how well the journal name represents a publication, ensuring a higher diversity of journals in the training set is a more effective approach to improving classification performance.
We provide our results for experiments on data partitioned with the constraint of distinct journals across datasets and without this constraint.

Choice of model configuration

As shown in Table 1, the resulting model should be able to work on different configurations of textual features. There are two possible approaches to achieve this. In the first approach, separate models could be built for each feature and combination of features. A meta-model would then determine, based on the textual data available for the instance to be classified, which of these models should be applied to achieve the best performance. In contrast, the second approach involves training a single

unified model on various combinations of publication textual features. This unified model is designed to handle any combination of the three textual features as input.
The second approach offers the advantage of being more compact and easier to use. However, it is assumed that the first approach might perform better on specific features since each model is exclusively trained on its corresponding configuration. In the results section, these two approaches will be compared, alongside the individual performance of each model.
The models will be evaluated using precision, recall, and the F1-score, which is the harmonic mean of precision and recall. While we have aimed to create a relatively balanced test set, perfect balance cannot be ensured in a multilabel scenario. As a result, we focus on macro scores (calculated as the average of class-wise metrics) rather than micro scores (calculated for the dataset as a whole). This ensures that performance is assessed at the level of individual disciplines, rather than being influenced by the potentially higher representation of certain disciplines in the dataset.

Part 2: Application to VABB-SHW

Translation

For this research, we opted to translate all non-English VABB-SHW publications into English to simplify the problem. We used the GPT-4o-mini model for this task. Although no studies have yet evaluated the quality of translation done by the GPT-4o-mini model, findings from the shared task in translation from the Workshop on Statistical Machine Translation (Kocmi et al., 2024) show promising results for its predecessor, GPT-4, positioning it as the top performing model for English-German translation quality (German is the language closest to Dutch from the list) based on human evaluation. Hendy et al., 2023 evaluated another one of its predecessors, GPT-3.5 (text-davinci-003), on translation tasks in comparison with other existing models and software. Some of the main conclusions are that translations produced by GPT are more fluent, achieving consistently lower perplexity and more non-monotonic, producing translations with longer range reordering. However, the authors have also noted that given that the models are not specialized in translation or trained on parallel texts in multiple languages, LLMs are less constrained in their faithfulness to the source text compared to translation-specialized models.
Nevertheless, we consider that for the task at hand a fluent, context-appropriate translation of the proposed text is sufficient to extract information regarding the discipline affiliation. Moreover, given previous comparisons of the GPT models on other tasks, we expect GPT-4o-mini to achieve superior results to its predecessors.

Evaluation of the classification

The methodology for evaluating the classification depends on the classification type. We combine automated testing with manual testing to estimate how reliable the database classification is at an individual publication level.

For evaluation of book classification, we use the existing classification based on international databases and manual classification at the level of book. In total, 53.01% of books already have a classification in the database .
A subset of the conference proceeding papers and the journal articles are classified manually by a member of our team with no prior access to the models' classification. The subset is selected based on previous classification experiments such that 0.10% of publications for each discipline are in the sample, but not less than five, in total 554 publications. The annotator has received a shuffled version of the data with no prior knowledge of how it has been selected.
A similar procedure is applied to a portion of the data for book chapters, with 0.30% of publications for each discipline included in the sample, again with a minimum of five publications per discipline, summing to 457 publications. This approach aims to preserve the supposed distribution of disciplines in the dataset, with a minimum representation for all.
Another part of the evaluation of book focuses on chapters with generic names, defined as instances where more than 15 book chapters share the same name. These chapters are expected to should be classified the same as the originating book and are excluded from the manual classification sample. The top 10 most frequent book chapter names are shown in Table 2. These names are typically variations of generic book sections (e.g., introductions, conclusions) or chapters about Belgium.

Table 2. Top 10 most frequent book chapter names.

Chapter title	*English translation*	*Count*
Introduction	Introduction	735
Inleiding	Introduction	235
Belgium	Belgium	205
Preface	Preface	148
Voorwoord	Foreword	125
Woord vooraf	Foreword	112
Foreword	Foreword	82
Conclusion	Conclusion	45
Préface	Preface	43
Ten geleide	Introduction	31

For this part of the analysis, since the test data partially reflects the discipline repartition in VABB-SHW for the specific publication type, we will focus on micro metrics (micro-precision, micro-recall, micro-F1). This approach aligns with our interest in evaluating the model's overall performance on the entire sample rather than its performance at the discipline level.

Results

Threshold analysis

As outlined in the methodology, the threshold selection is done in two steps: first, candidate thresholds are identified based on the distribution of labels, and second, the final threshold is selected for the model based on classification results with abstracts.

Table 3 presents the distribution of the number of labels for thresholds ranging from 0.0 (a discipline is assigned to a publication if any referenced publication is classified into that discipline in WoS) to 1.0 (a discipline is assigned only if all referenced publications are classified into that discipline in WoS). The thresholds 0.25 to 0.55 respect the constraint of having more than 90% of the publications into one to three labels, thus they are retained for further testing.

Table 3. Distribution of the number of disciplines per publication across different thresholds.

Threshold	*0 labels*	*1 label*	*2 labels*	*3 labels*	*4 labels*	*5+ labels*	*Share with 1-3 labels*
0.0	1 932	719 570	918 916	1 151 588	1 220 675	3 960 541	34.99
0.05	1 935	960 321	1 278 720	1 717 112	1 504 943	2 510 191	49.62
0.1	1 954	1 396 688	1 851 380	2 074 124	1 408 808	1 240 268	66.75
0.15	2 088	1 833 008	2 325 294	2 102 251	1 123 273	587 308	78.52
0.2	2 896	2 363 844	2 746 408	1 886 007	750 042	224 025	87.75
0.25	6 204	2 928 016	2 987 144	1 520 333	448 764	82 761	**93.26**
0.3	14 186	3 472 914	3 025 206	1 153 258	270 289	37 369	**95.96**
0.35	43 491	4 121 897	2 858 108	787 419	148 342	13 965	**97.42**
0.4	110 930	4 789 151	2 486 922	497 389	82 194	6 636	**97.49**
0.45	217 847	5 273 439	2 087 337	337 452	52 578	4 569	**96.55**
0.5	481 774	5 767 008	1 517 250	180 793	25 125	1 272	**93.63**
0.55	692 572	5 894 650	1 233 191	133 630	18 125	1 054	**91.07**
0.6	1 060 689	5 922 292	893 183	84 780	11 544	734	86.54
0.65	1 412 112	5 804 604	686 108	61 275	8 475	648	82.17
0.7	1 869 712	5 579 591	480 314	37 771	5 398	436	76.48
0.75	2 361 706	5 248 435	334 441	24 724	3 564	352	70.33
0.8	2 811 153	4 895 797	245 374	18 007	2 577	314	64.71
0.85	3 234 406	4 533 138	188 938	14 419	2 021	300	59.41
0.9	3 723 689	4 095 108	140 776	11 678	1 684	287	53.27
0.95	4 172 587	3 672 019	116 039	10 716	1 577	284	47.64
1.0	5 282 625	2 669 322	21 248	27	0	0	33.75

Table 4 shows the macro scores for each threshold on the threshold's test data. With the exception of 0.25, the values tend to peak at 0.5 and then start going down. Threshold 0.25 is notable for its higher representation for Other Humanities and Health Biotechnology, which are otherwise significantly underrepresented for the other thresholds and often with a F1-score of 0. Excluding these two disciplines would result in similar values between 0.25 and 0.5.
For the next part, the results with the 0.5 threshold are presented. However, for the classification analysis of VABB-SHW, the results with both models are tested to reverify which is the more accurate model.

Table 4. Classification results for different thresholds.

Threshold	0.25	0.30	0.35	0.40	0.45	0.50	0.55
Macro recall	82.94%	75.90%	75.16%	76.99%	76.09%	76.46%	76.92%
Macro precision	73.34%	71.52%	71.73%	71.28%	71.33%	72.81%	71.28%
Macro F1-score	**76.48%**	73.33%	73.15%	73.78%	73.38%	**74.27%**	73.72%

Results for WoS data

First, we evaluate the impact of journal names (channel title) on the quality of the classification. This is done by using distinct journals for the train, validation and test dataset as explained in the Methodology section. The results are presented in
Table 5.
The journal name is a poor predictor of the discipline of the publication (7.44 % macro F1-score) and the increase in the quality of prediction when the journal name is added to the article title is insignificant (59.80% macro F1-score for title only and 60.38% for title and journal title). There is in fact a decrease when the journal name is added to the abstract (66.92% macro F1-score for abstract only and 65.95% with abstract and journal title). This result is not surprising given that when only the journal name is used as a feature to predict publication classification, the same journal can have different classifications assigned in the train dataset as the entity classified is the publication, not the journal.
Therefore, as mentioned in the Methodology section, to increase the variety of publications in a discipline, we have decided that for final model selection we ignore this restriction and allow publications from the same model to be present in the train, validation and test database. Modelled like this, the problem is a more realistic representation of the general classification problem studied in this research that should not exclude the benefit given by the presence of the journal in the train database.

Table 5. Classification results for train, validation and test datasets containing distinct journals.

Model data	Macro precision	Macro recall	Macro F1
Abstract only	72.72%	63.91%	66.92%
Channel title + Abstract	71.33%	63.32%	65.95%
Title only	67.79%	55.74%	59.80%
Channel title + Title	68.30%	56.24%	60.38%
Channel title	32.17%	4.47%	7.44%

In general, the discrepancy between the results of the predictions when allowing (Table 6) publications from the same journals in train, validation and test set – whether or not the journal name is used as a feature – compared to when the publications in the three sets come from distinct journals (

Table 5) point towards journal specialization resulting in publications from the journals in the train set being a worse representation of the ones in the test and validation sets when they come from other journals from that discipline.

Table 6. Classification results for the train, validation and test datasets selected with no restriction at the level of journal. (-) marks the models that would not be used for the final classification.

Model data	Macro precision	Macro recall	Macro-F1	Rank
Abstract only	76.94%	72.03%	74.12%	5 (-)
Title only	72.79%	62.87%	66.94%	6
Title + Abstract	76.94%	73.07%	74.77%	3
Channel title + Abstract	77.97%	75.67%	76.63%	2 (-)
Channel title + Title	77.11%	72.28%	74.33%	4
Channel title + Title + Abstract	78.38%	75.81%	76.91%	1
Combined	77.08%	71.78%	74.04%	

Table 6 shows the results of the classification when no restrictions are applied on the channel of the classified article. The table includes results for individual features, combinations of features, and a combined model. The combined model is trained on the merged training data from the other experiments, meaning it includes examples with only abstracts, examples with both abstract and title, examples with only the title, and so on.

A meta-model would need to address 4 possible combinations of features in the VABB-SHW dataset: all the features are available, only the title and the channel title are available, only the title and the abstract are available, and only the title is available. The results in Table 6 indicates that the model that is trained on all the available features should be used for all the scenarios.

Since the training, validation and test datasets for all the previous models consists of the same articles, but with different textual features put forward, the combined data is six times larger than the individual datasets. It includes the same articles six times, but represented by distinct features or combinations of features. The next experiment aims to determine whether building a single model capable of classifying data with different structures results in any loss of prediction quality.

To properly evaluate the combined model, its performance must be tested on the individual test datasets to assess whether it underperforms or overperforms compared to models specialized for specific features or feature combinations. Table 7 presents these results, showing that variations in the F1 score are not significant to conclude that the combined model performs better or worse than the models specialized on a feature or a group of features.

Based on these findings, we focus our further analysis on the combined model, as it can be applied to the VABB-SHW dataset as a whole, even in cases where certain features are missing.

Table 7. Classification results for the combined model when tested on individual features and feature combinations.

Test data	**Macro-F1**	**Comparison Macro**
Abstract only	74.21%	+ 0.09%
Title only	67.24%	+ 0.30%
Title + Abstract	74.49%	- 0.28%
Channel title + Abstract	76.51%	- 0.12%
Channel title + Title	74.14%	-0.19%
Channel title + Title + Abstract	76.56%	-0.35%

Results for VABB-SHW

Table 8 shows the results for all the available labelled datasets originating from the VABB-SHW dataset.

Table 8. Classification results for the available labelled VABB-SHW datasets.

		Threshold 0.5			*Threshold 0.25*		
Test set	*Nb. Pub.*	*Micro Precis.*	*Micro Recall*	*Micro F1-score*	*Micro Precis.*	*Micro Recall*	*Micro F1-score*
Manual journal articles and conference proceedings	554	50.25%	58.25%	53.96%	42.65%	65.26%	51.59%
Manual book chapters	457	56.31%	60.11%	58.15%	51.27%	67.91%	58.43%
Book chapters with generic names	339	51.92%	51.92%	51.92%	47.82%	57.14%	52.07%
Books (from previous classification)	14 916	55.14%	55.41%	55.27%	51.63%	61.98%	56.33%
Total	16 266	54.90%	55.59%	55.25%	51.11%	62.19%	56.11%

For the manual classification, and book classification datasets, the results consistently achieve an F1-score of 54–58%. However, book chapters with generic names score lower, likely due to the noise introduced by the chapter name and the overall lack of sufficient textual data for accurate classification. When comparing the 0.25 threshold with the 0.5 threshold, the former gains in recall but loses in precision. This is because the 0.25 threshold predicts a larger number of labels.

To further understand the classification results, Table 9 presents the outcomes for the top 10 most represented disciplines in the total VABB-SHW test dataset (the combination of all test datasets for VABB-SHW), including the results of the combined model on the WoS test data. Disciplines that are easily identified in VABB-SHW, such as Law and Language and Linguistics also achieve good results on WoS data. In contrast, History, Art, and Sociology underperform on both test datasets, with Sociology proving particularly challenging for the model to classify accurately.

Economics and Business, Philosophy and Ethics, and Political Science are notable cases. While these disciplines perform well on WoS data, they underperform on VABB-SHW data. This discrepancy may indicate that the training data does not adequately represent these disciplines as they appear in VABB-SHW. Alternatively,

given that the book dataset is the largest in the test datasets, the definition of these disciplines, as inferred from journal articles, may not translate well to other publication types.

To investigate this further, Table 10 presents the results for these disciplines in the individual test datasets. The findings for manually annotated datasets outperform those for the total test set. Additionally, differences between the dataset containing journal articles and conference proceedings and the one with book chapters suggest that publication types significantly impact classification performance. Furthermore, the differences between the manually annotated datasets and the rest may also be, at least in part, due to variations in the annotation methodology across datasets.

Table 9. Classification results for top 10 disciplines based on the frequency in the total test set for VABB-SHW, threshold 0.5.

Discipline	*# instances in combined test set*	*Precision*	*Recall*	*F1-score*	*F1-score for WoS data*
Law	3 532	89.47%	79.16%	84.00%	89.34%
Literature	1 673	52.35%	67.12%	58.83%	67.60%
History	1 524	37.06%	54.00%	43.95%	62.24%
Sociology	1 395	52.60%	21.79%	30.82%	48.70%
Languages and linguistics	1 344	85.33%	61.021%	71.15%	83.75%
Economics and business	1 262	67.73%	45.56%	54.48%	74.26%
Art	1 217	58.12%	55.88%	56.98%	66.61%
Religion	1 199	78.59%	48.37%	59.89%	70.95%
Political Science	1 041	43.35%	56.00%	48.87%	75.53%
Philosophy and ethics	914	46.64%	56.24%	50.99%	77.54%

Table 10. Classification results for Economics and business, Political Science and Philosophy and ethics across different VABB-SHW test datasets.

Dataset	*Metric*	*Economics and business*	*Political Science*	*Philosophy and ethics*
Manual journal articles and conference proceedings	*# instances*	61	39	17
	F1-score	66.10%	54.55%	57.89%
Manual book chapters	*# instances*	57	43	25

	F1-score	58.59%	55.56%	74.51%
Book chapters with generic names	*# instances*	31	13	25
	F1-score	41.86%	43.90%	51.52%
Books (from previous classification)	*# instances*	1 113	946	847
	F1-score	53.81%	48.57%	50.19%
WoS data	*F1-score*	74.26%	75.53%	77.54%
Total VABB-SHW test data	*# instances*	1 262	1 041	914
	F1-score	54.48%	48.87%	50.99%

Conclusion

This research presents a methodology for classifying publications from local databases based solely on textual information. We divided the analysis into two parts: one focused on building the model, and the other on applying it to classify the publications included in the Flemish database for Social Sciences and Humanities (VABB-SHW).

In the first part, we investigated which ground truth strategy best represents the data while maintaining an optimal number of disciplines per publication. The range for the optimal threshold was narrowed to 0.25–0.55. Based on classification results across various thresholds, we selected the 0.5 threshold for further analysis of how to address the availability of different textual features. However, given the promising results of the 0.25 threshold, it was also considered for the VABB-SHW data.

Additionally, we evaluated two strategies to address the potential lack of certain textual features in the VABB-SHW data. The first strategy involved using a meta-model that selects among feature-specific models, while the second proposed a single model trained on various textual features and feature combinations to handle varied input. The results showed similar performance for both strategies, and we opted for the combined model due to its ease of application.

When analyzing the classification results on VABB-SHW, we observed significantly worse performance on the VABB-SHW test dataset compared to the WoS test dataset. One identified factor contributing to this discrepancy is the publication type.

Discussions and Limitations

Other factors, such as the availability of textual features, translation errors, local terminology, and specific topics, may also contribute to the observed discrepancies between the results for VABB-SHW and WoS. We have currently not yet explored these aspects in detail but this could provide valuable insights in future research.

While this research has provided overall metrics for classification performance, it has not qualitatively analysed the nature of the classification errors. Future work

could involve examining disciplines that are frequently misclassified and investigating whether errors stem from true misclassification or differences in interpretation. Given the absence of an incontestable ground truth for discipline classification and the fact that some publications lie at the intersection of multiple disciplines, some errors may involve such borderline cases.

This study has certain limitations that should be considered while interpreting the results. First, the methodology relies on the classification of references in WoS to infer the final ground truth. Consequently, the model is trained to predict the disciplines associated with the journals most commonly cited by the publication, using this as a proxy for the discipline of its content.

Secondly, we assume that the selected classification scheme accurately represents the underlying structure of the data and that the model can effectively learn to distinguish each discipline based on the provided examples. However, this assumption has not yet been empirically tested, as the classification scheme was chosen based on its alignment with other types of classification in the database rather than its specific suitability for the data.

Thirdly, the evaluation was conducted on a small sample of VABB-SHW publications, which may not fully capture the diversity of the dataset, especially for journal articles, conference proceedings, and book chapters. Expanding this sample in future research would provide a more comprehensive understanding.

Fourthly, the data for non-SSH disciplines in VABB-SHW consists of publications (co-)authored by Flemish researchers from SSH departments. As a result, this content may deviate slightly from the typical literature in those fields. Exploring this aspect further could shed light on its potential impact.

Finally, the study assumes that disciplines are static over time, which has been shown by previous research (Manning, 2020; Zhou et al., 2022) to be an oversimplification. While the time dimension was not explicitly accounted for in this analysis, its potential influence represents an interesting direction for future exploration.

Acknowledgments

We want to thank our colleague, Eline Vandewalle, for the manual annotation of the data.

References

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingrasb, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342. https://doi.org/10.1007/s11192-006-0115-z

Arhiliuc, C., Guns, R., Daelemans, W., & Engels, T. C. E. (2024). Journal article classification using abstracts: A comparison of classical and transformer-based machine learning methods. *Scientometrics*. https://doi.org/10.1007/s11192-024-05217-7

Aviv-Reuven, S., & Rosenfeld, A. (2023). A logical set theory approach to journal subject classification analysis: Intra-system irregularities and inter-system discrepancies in Web of Science and Scopus. *Scientometrics*, *128*(1), 157–175. https://doi.org/10.1007/s11192-022-04576-3

Guns, R., Sīle, L., Eykens, J., Verleysen, F. T., & Engels, T. C. E. (2018). A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics*, *116*(2), 1093–1111. https://doi.org/10.1007/s11192-018-2775-x

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation* (arXiv:2302.09210). arXiv. https://doi.org/10.48550/arXiv.2302.09210

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., … Zouhar, V. (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 1–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Manning, P. (2020). Disciplines and Their Evolution. In P. Manning (Ed.), *Methods for Human History: Studying Social, Cultural, and Biological Evolution* (pp. 83–90). Springer International Publishing. https://doi.org/10.1007/978-3-030-53882-8_8

Milojević, S. (2020). Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, *1*(1), 183–206. https://doi.org/10.1162/qss_a_00014

OECD. (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2015_9789264239012-en

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2017). A comparison of the Web of Science and publication-level classification systems of science. *Journal of Informetrics*, *11*(1), 32–45. https://doi.org/10.1016/j.joi.2016.10.007

Shen, S., Liu, J., Lin, L., Huang, Y., Zhang, L., Liu, C., Feng, Y., & Wang, D. (2022). SsciBERT: A pre-trained language model for social science texts. *Scientometrics*. https://doi.org/10.1007/s11192-022-04602-4

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. (2017). *European Databases and Repositories for Social Sciences and Humanities Research Output*. https://doi.org/10.6084/M9.FIGSHARE.5172322

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E., Arefiev, P., Dušková, M., Faurbæk, L., Holl, A., Kulczycki, E., Macan, B., Nelhans, G., Petr, M., Pisk, M., Soós, S., Stojanovski, J., Stone, A., Šušol, J., & Teitelbaum, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, *27*(4), 310–322. https://doi.org/10.1093/reseval/rvy016

Singh, P., Piryani, R., Singh, V. K., & Pinto, D. (2020). Revisiting subject classification in academic databases: A comparison of the classification accuracy of Web of Science, Scopus & Dimensions. *Journal of Intelligent & Fuzzy Systems*, *39*(2), 2471–2476. https://doi.org/10.3233/JIFS-179906

Sivertsen, G. (2016). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, *107*(2), 357–368. https://doi.org/10.1007/s11192-016-1845-1

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, *91*(2), 567–575. https://doi.org/10.1007/s11192-011-0615-3

Verleysen, F., Ghesquière, P., & Engels, T. (2014). *The objectives, design and selection process of the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW).*

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Zhou, H., Guns, R., & Engels, T. C. E. (2022). Are social sciences becoming more interdisciplinary? Evidence from publications 1960–2014. *Journal of the Association for Information Science and Technology*, *73*(9), 1201–1221. https://doi.org/10.1002/asi.24627

https://doi.org/10.51408/issi2025_091

The Effects of Research Evaluation: Do Researchers' Perceptions Align with Evidence?

Giovanni Abramo,[1] Ciriaco Andrea D'Angelo,[2] Emanuela Reale,[3] Antonio Zinilli[4]

[1]*giovanni.abramo@unimercatorum.it*
Universitas Mercatorum, Laboratory for Studies in Research Evaluation
Piazza Mattei 10, 00186 Roma (Italy)

[2]*dangelo@dii.uniroma2.it*
University of Rome "Tor Vergata", Dept of Engineering and Management
Via del Politecnico 1, 00133 Roma (Italy)

[3]*emanuela.reale@cnr.it*, [4]*antonio.zinilli@cnr.it*
CNR-IRCrES, Consiglio Nazionale delle Ricerche, Research Institute on Sustainable Economic Growth,
Via dei Taurini 19, Roma (Italy)

Abstract

This study examines the alignment between researchers' perceptions of the Italian Scientific Habilitation (ASN) and the bibliometric evidence regarding its impact on scientific productivity in STEMM disciplines. The ASN, introduced in 2012, serves as a key evaluation tool for academic promotions in Italy, aiming to enhance research productivity and quality, and contrast favoritism. Employing a mixed-methods approach, we compare survey data from academics with bibliometric analyses of publication output over two five-year periods (2008–2012 and 2013–2017). The findings reveal significant misalignments: while bibliometric evidence indicates measurable productivity increases following the introduction of the ASN, survey responses suggest that many researchers perceive little to no impact.

The divergence between perception and evidence varies across demographic and disciplinary contexts. Younger researchers and early-career academics report stronger perceived and measurable productivity increases, reflecting their reliance on the ASN for career progression. In contrast, older researchers show measurable gains in bibliometric analyses but often do not attribute these improvements to the evaluation system. Disciplinary differences also emerge: fields such as Medicine and Engineering exhibit high productivity gains in both perception and evidence, while disciplines like Physics and Mathematics demonstrate significant bibliometric increases but low perceived impact.

This mismatch carries critical implications for research evaluation practices. For researchers, it highlights a potential erosion of trust in evaluation systems, particularly among those who feel their contributions are undervalued. This discontent may lead to disengagement or counterproductive behaviors, such as prioritizing short-term outputs or engaging in unethical practices like self-citation, citation networks, or salami-slicing publications. For institutions, the findings underscore the need to tailor evaluation practices to accommodate disciplinary differences and to recognize diverse contributions beyond publications, such as teaching and societal impact.

At the policy level, the study advocates for a more inclusive and transparent evaluation framework. Recommendations include integrating qualitative assessments with bibliometric metrics, developing discipline-specific evaluation criteria, and addressing disparities in gender, geography, and institutional resources. Efforts to enhance transparency and communication in evaluation systems could bridge the gap between perception and evidence, fostering greater trust and legitimacy.

Despite its contributions, the study has limitations. The survey data captures subjective perceptions that may be influenced by personal biases, while bibliometric analyses rely on productivity proxies

that overlook qualitative aspects of research. Future research should employ longitudinal and qualitative methods to explore the underlying causes of misalignment and its impact on academic behavior.
By addressing the roots of the mismatch between perception and evidence, this study provides actionable insights for designing evaluation systems that align with academic values, promote equity, and incentivize long-term innovation.

Introduction

Research evaluation tools have become indispensable for assessing the pursuit of research policy goals and strategic objectives. They focus mainly on key dimensions of research performance such as productivity, quality, and impact of academic work (de Diego et al., 2024). These systems influence various decisions, create specific individual incentives, and stimulate organisational and management changes. However, their implementation has sparked an ongoing debate about their unintended consequences and the extent to which they align with the broader goals of scientific inquiry (de Rijcke et al., 2016). Central to this debate is the question of whether researchers' perceptions of these systems match empirical evidence regarding their effects. Misalignments between perception and evidence can distort academic priorities, undermine equity, and inhibit the cultivation of diverse intellectual landscapes.

In this article, we explore the complex dynamics between perception and evidence in the context of research evaluation. In particular, we intend to contrast outcomes related to the changes in research productivity (increase or decrease of productivity) from a survey-based study with those arising from bibliometric pictures, taking Italy as a field of observation since the country was recently interested in the heavy introduction of research assessment. In this work, by research productivity, we mean the publications produced by a researcher over a given time period, as this is the most widely accepted definition in academia and, therefore, suitable for use in a survey.[1] In particular, we concentrate on a research evaluation exercise named the Italian Scientific Habilitation (ASN).[2] Introduced for the first time in 2012, it enables habilitated individuals to be selected for positions of Associate professors and Full professors in Italian universities. Therefore, the evaluation exercise analysed in this paper is strongly related to the academic career. Our investigation focuses exclusively on STEMM fields, which have distinct publication practices and research evaluation dynamics compared to other disciplines. Above all, they are particularly well-suited for bibliometric evaluation.

The development and proliferation of research evaluation metrics have transformed academic ecosystems. Metrics such as the journal impact factor (Garfield, 1972), citation counts (Bornmann & Daniel, 2008), and the h-index (Hirsch, 2005) were initially designed to complement qualitative assessments of research quality. According to a few scholars, their widespread adoption has led to an over-reliance on quantitative measures, often reducing complex scholarly contributions to narrow,

[1] For a more detailed definition of research productivity, we refer the reader to Abramo and D'Angelo (2014).
[2] https://abilitazione.mur.gov.it/public/index.php?lang=eng.

one-dimensional scores. This “metric fixation” (Muller, 2018) has contributed to several well-documented issues, including the reinforcement of existing inequalities, a bias toward mainstream disciplines, and the undervaluation of less measurable dimensions of academic work, such as teaching and mentorship (McKiernan et al., 2016). Other scholars hold that the problem with metrics is that they are applied by individuals without professional expertise, while evaluative scientometricians know well in which circumstances to adopt scientometrics and in which to recur to other methods (Abramo, 2024; Ioannidis & Maniadis, 2023).

Researchers’ perceptions of these evaluation systems often reflect frustration with their perceived rigidity, bias, and opacity. Surveys indicate that many researchers feel pressured to prioritize short-term outputs, such as publishing in high-impact journals, over long-term goals, such as fostering innovation or addressing societal challenges (Nicholas et al., 2017; Fire & Guestrin, 2019). Moreover, qualitative studies suggest that evaluation systems can create misaligned incentives, encouraging practices such as salami-slicing publications or favoring “safe” research over more exploratory or interdisciplinary work (Sahel, 2011; Brembs et al., 2013). While these perceptions are widely reported, empirical evidence presents a more nuanced picture of the effects of evaluation systems, highlighting both their benefits and drawbacks (Abramo & D’Angelo, 2021; Seeber et al., 2019).

Empirical studies reveal that evaluation metrics can effectively identify high-impact research and facilitate comparisons across disciplines and institutions (Waltman, 2016). However, they also underscore significant limitations. For instance, citation-based metrics are heavily influenced by field-specific publication practices, with some disciplines inherently generating fewer citations than others (Moed, 2005). Additionally, gender and geographic disparities persist, with women and researchers from the Global South often receiving less recognition and fewer citations, even when their work is of comparable quality (Larivière et al., 2013). These findings challenge the assumption that using metrics for research evaluation is neutral or universally applicable, suggesting that researchers' perceptions of bias may be well-founded.

The mismatch between perception and evidence in research evaluation has profound implications. When researchers perceive evaluation systems as unfair or misaligned with academic values, it can erode trust, reduce motivation, and lead to gaming behaviors that undermine the integrity of the scientific process (Smaldino & McElreath, 2016). Conversely, efforts to address this misalignment—such as initiatives promoting responsible research assessment (DORA, 2012; Hicks et al., 2015) and the use of narrative CVs (Moher et al., 2022)—have shown promise in fostering more equitable and holistic evaluation practices, which may mitigate these negative effects.

This paper aims to examine the misalignment between the perceptions of researchers and empirical evidence in research evaluation, specifically in relation to the impact of the ASN on scientific productivity, understood as the increase in scientific publications since its introduction.

To achieve this, within the context delineated above, the paper addresses the following question: “Is there a misalignment between researchers' perceptions and

empirical evidence regarding the effects of research evaluation on productivity, when controlling for individual and contextual factors?"

The findings of the study can help formulate actionable strategies for bridging the gap between perception and evidence. By integrating insights from bibliometric research, sociology of science, and policy studies, we aim to provide a comprehensive understanding of how research evaluation systems shape academic behavior. Addressing the misalignment between perception and evidence is not merely a matter of improving metrics or processes; it is essential for restoring trust, promoting inclusivity, and ensuring that research evaluation serves its intended purpose of advancing knowledge and societal well-being. Through this lens, we aim to contribute to the ongoing dialogue on building evaluation systems that align with the values and realities of the research community.

The paper is organized as follows. In the next section, we will illustrate the methodological issues of the two proposed analyses and, in the following, the main results of the analyses and their comparison. The concluding section summarizes the main findings and illustrates the authors' considerations about implications and future developments.

Methods

This paper uses both survey and bibliometric analyses to explore the factors influencing the impact of the ASN on scientific productivity in STEMM disciplines. The analyses share a consistent framework of independent variables, ensuring comparability between the subjective perceptions captured in the survey and the outcomes derived from bibliometric data. Using the same set of independent variables, we aim to provide a comprehensive understanding of how individual, institutional, geographic, and disciplinary factors shape the perception of the ASN and its measurable effects.

The factors or independent variables included in the analyses are:

- Individual factors: Gender (male vs female) and age groups (<35, 35–44, 45–54, 55–65, and >65, with the oldest group serving as the reference category).
- Institutional size: universities are categorized as large- (reference category), medium-sized, and small-sized.
- Geographic location: Regions are categorized as North, Centre, and South (reference category).
- Disciplinary areas: The analysis includes 10 (STEMM)[3] of the 14 Italian university disciplinary areas (CUN), with Physics (CUN 2) serving as the baseline category.[4] We exclude from the analysis the areas of social sciences

[3] 1 - Mathematics and computer science, 2 - Physics, 3 - Chemistry, 4 - Earth sciences, 5 - Biology, 6 - Medicine, 7 – Agricultural and veterinary sciences, 8 - Civil engineering, 9 - Industrial and information engineering, 10 - Psychology.

[4] Physics is chosen as the baseline category for two reasons: i) Physics is a well-established field with relatively standardized research and publication practices. It provides a consistent benchmark for comparison with other disciplines that may have more diverse or variable practices; ii) Physics is known for its high volume of publications and collaborations, often within large international research

and arts and humanities, due to the limited coverage in bibliographic repertories of the research output in these areas (Mongeon & Paul-Hus, 2016; Archambault et al., 2006). For the area "Historical, Philosophical, Educational, and Psychological Sciences," only the subarea of Psychology is included in the analysis, as eligible for bibliometric analysis.

In the survey analysis, we assess the perceptions of researchers on whether the ASN has influenced their scientific productivity. Respondents were asked to consider the last ten years of their career, which means from date back until 2012 when ASN was introduced in Italy. In contrast, the bibliometric analysis measures actual changes in productivity, using a binary outcome variable indicating whether there was an increase in publication output between two five-year periods (2008–2012 and 2013–2017), i.e. after the introduction of the ASN. By combining these approaches, we can compare the perceptions with evidence, identifying both areas of alignment and divergence.

The survey

The data for this study were collected through a national survey conducted in Italy between 2020 and 2021. The survey used a structured questionnaire administered to a probabilistic sample of academics from Italian universities in the disciplinary areas under observation. The survey collected information on the effects of the ASN, focusing on individual adaptation or response, as well as respondent characteristics (e.g., gender, age, academic position) and institutional contexts (e.g., university size). For geographic distribution, the adopted classification is into three main macro-areas: North, Centre, and South. Regarding academic ranks, the study included five positions introduced by the Gelmini Law (L. 240/2010): researcher, type A researcher (RTD-A), type B researcher (RTD-B), associate professor, and full professor. In the following Table 1, a detailed breakdown of the survey dataset is provided.

Table 1. Breakdown of the dataset (822 professors) by personal and contextual variables.

Variable	Level	Share
Gender	F	36.1%
	M	63.9%
Age	Less than 35	1.7%
	35-44	19.6%
	45-54	34.1%
	55-65	34.5%
	Over 65	10.1%
Univ. size	Big	47.5%
	Medium	33.8%
	Small	18.7%
Univ. location	South	26.3%
	Center	26.8%
	North	46.9%

teams. Its citation practices and publishing norms are relatively well-aligned with bibliometric indicators commonly used in evaluation systems like the ASN.

We applied a logit model to analyze the likelihood of response and to identify the factors influencing respondents' perceptions of the ASN's impact on their productivity. This approach allows us to derive a regression equation capable of predicting the category each academic falls into, based on the explanatory variables. The dependent variable in this study was constructed using the survey question: "*In the past ten years, to what extent have the following factors influenced the quantity of your publications?"*

This question captures a range of influences on scientific productivity, including, but not limited to, the ASN. Factors considered include, for instance, the need to align with ASN requirements, gaining a competitive edge in securing research funding, participating in national or international research projects, and increasing academic visibility. These additional factors provide a comprehensive view of the various motivations and external pressures that may impact the quantity of publications.

To isolate the effect of the ASN from other factors, we focused specifically on respondents who reported an increase in productivity and explicitly attributed this change to the ASN. By narrowing the analysis to this subgroup, we were able to disentangle the impact of the ASN from other influences, allowing for a more targeted assessment of its role in shaping research output. This approach ensures that our findings reflect the specific contribution of the ASN, separate from broader or overlapping factors. The analysis focuses on 822 respondents belonging to STEMM scientific areas.

The bibliometric analysis

Our dataset comprises 26,217 professors (assistant, associate, or full) from Italian universities, who held tenured positions in STEMM fields continuously from 2008 to 2017. Table 2 shows their distribution by academic field and rank, based on data as of December 31, 2012, i.e. around the time the ASN was introduced for the first time in Italian academia. Table 3 summarizes the relative frequencies of personal variables (gender and age) and contextual variables (size and location of the university of affiliation).

Table 2. Dataset of the bibliometric analysis. Breakdown by field and academic rank.

Field*	Assistant prof.	Associate prof.	Full prof.	Total
1 – MATH	926 (38.5%)	794 (33.0%)	687 (28.5%)	2407 (9.2%)
2 – PHYS	620 (38.5%)	623 (38.6%)	369 (22.9%)	1612 (6.1%)
3 – CHEM	1024 (45.9%)	763 (34.2%)	446 (20.0%)	2233 (8.5%)
4 – EARTH	354 (44.6%)	286 (36.1%)	153 (19.3%)	793 (3.0%)
5 – BIOL	1725 (48.0%)	1070 (29.8%)	796 (22.2%)	3591 (13.7%)
6 – MED	3514 (49.0%)	2193 (30.6%)	1461 (20.4%)	7168 (27.3%)
7 – AGRVET	1034 (42.9%)	788 (32.7%)	587 (24.4%)	2409 (9.2%)
8 – CIVENG	435 (36.6%)	428 (36.0%)	326 (27.4%)	1189 (4.5%)
9 – INDENG	1422 (35.9%)	1377 (34.7%)	1167 (29.4%)	3966 (15.1%)
11 – PSYCH	353 (41.6%)	271 (31.9%)	225 (26.5%)	849 (3.2%)
Total	11407 (43.5%)	8593 (32.8%)	6217 (23.7%)	26217

* *1-Mathematics and computer science, 2-Physics, 3-Chemistry, 4-Earth sciences, 5-Biology, 6-Medicine, 7-Agricultural and veterinary sciences, 8-Civil engineering, 9-Industrial and information engineering, 10-Psychology.*

Table 3. Breakdown of the dataset (26.217 professors) by personal and context variables.

Variable	Level	Share
Gender	F	33.1%
	M	66.9%
Age	Less than 35	0.5%
	35-44	22.6%
	45-54	42.0%
	55-65	34.5%
	Over 65	0.4%
Univ. size	Big	65.0%
	Medium	34.1%
	Small	0.9%
Univ. location	South	28.9%
	Center	27.7%
	North	43.4%

All variables were extracted from the database of Italian professors maintained by the Minister of University and Research (MUR).[5]

For setting the bibliometric dataset, we used the author name disambiguation algorithm developed by D'Angelo, Giuffrida, and Abramo (2011), based on the coupling of the publications extracted from the Web of Science *core collection* by Clarivate Analytics and the MUR database. This algorithm assigns a WoS publication (articles, reviews, letters, and conference proceedings only) to a given professor if the latter:

- Has a name matching one of the authors in the publication byline;
- Is affiliated with one of the recognized universities listed in the publication's author addresses;
- Is associated with a discipline that aligns with the subject category (SC) of the publication;
- Was on staff as of December 31 of the year preceding the publication year.

Once we have assigned to each professor in the dataset the publications he/she has authored, we calculate two indicators, namely output (O) and fractional output (FO). The first is the simple count of the authored publications; the second is the fractional count, whereby we sum up the fractional contribution of the author to its publications, i.e., for each publication, the reciprocal of the number of co-authors

[5] For each professor this database provides information on their name and surname, gender, affiliation, discipline, field and academic rank, at close of each year. http://cercauniversita.cineca.it/php5/docenti/cerca.php, last access on 30 January 2025.

and, for publications in life science, also the co-authorship type (intramuros vs extramuros) and the position in the byline.[6]
Finally, we measure the effect of ASN in binary terms, i.e. through a dummy variable, taking the value 1 if the indicator (O or FO) measured in 2013-2017 is greater than the value measured in the previous five-year period (2008-2012); 0 otherwise.

Results

Descriptive statistics

In this section, we present the descriptive statistics derived from both the survey and bibliometric data, focusing on the perceived and measured effects of the ASN on scientific productivity. The survey data captures the perceptions of researchers regarding the effects of the ASN, while the bibliometric data reflects actual changes in productivity between two five-year periods (2008–2012 and 2013–2017). By examining the distribution of the ASN effect across gender, age groups, university size, geographical areas, and CUN disciplinary areas, we aim to highlight the alignment and discrepancies between perceived and measurable impacts of this evaluation tool in STEMM disciplines.
The proportions of researchers reporting an "ASN effect" versus "No ASN effect" from the survey differ from evidence revealed by the bibliometric analysis. The majority of respondents (65%) indicate "No ASN effect" on productivity. This suggests that most researchers perceive their productivity as not being significantly influenced by the ASN. In contrast, the bibliometric analysis shows the opposite pattern, with the "ASN effect" representing the majority (62%), reflecting measurable increases in productivity attributed to the ASN.
The following Figure 1 presents the distribution of the ASN effect ("ASN effect") and no ASN effect ("No ASN effect") by gender, based on survey and bibliometric data.

[6] For the life sciences, widespread practice in Italy is for the authors to indicate the various contributions to the published research by the order of the names in the listing of the authors. For the life science SCs publications, we give different weights to each co-author according to their position in the list of authors and the character of the co-authorship (intra-mural or extra-mural) as suggested in Abramo, D'Angelo and Rosati (2013). If the first and last authors belong to the same university, 40% of contribution is assigned to each of them, the remaining 20% is divided among all other authors. If the first two and last two authors belong to different universities, 30% of contribution is assigned to the first and last authors, 15% of the citation is attributed to the second and last authors but one, the remaining 10% is divided among all others.

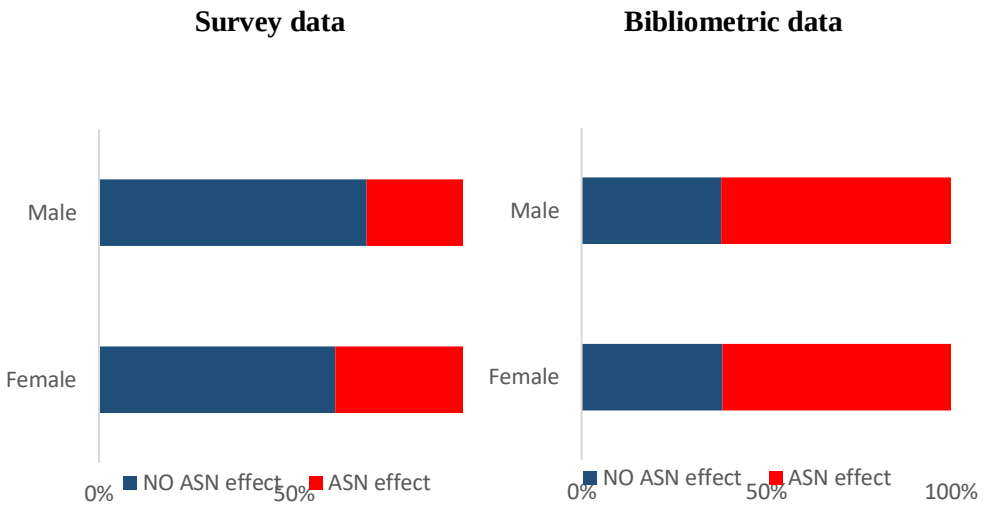


Figure 1. ASN effect by gender.

In the survey data, for both males and females, the majority report that the ASN has no effect on their scientific productivity. The bibliometric data, however, reveals a different pattern. In this case, the ASN effect appears to be more significant for both males and females, indicating a measurable increase in their scientific productivity. Figure 2 shows the distribution of the ASN effect ("ASN effect") and no ASN effect ("No ASN effect") for both survey and bibliometric data across different age groups: <35, 35-44, 45-54, 55-65, >65.

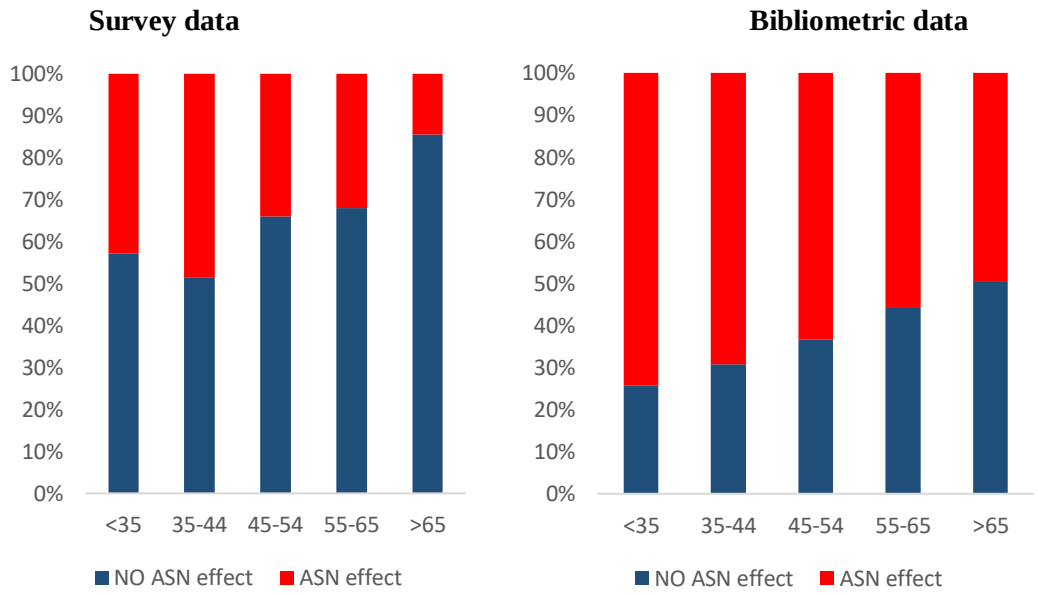


Figure 2. ASN effect by age.

In the survey-based chart, younger academics (<35 and 35–44 age groups) report a higher proportion of the "ASN effect" (red), indicating that these groups perceive a stronger impact of the ASN on their scientific productivity. The proportion of the "ASN effect" decreases progressively with age, becoming particularly small in the >65 group, where the "No ASN effect" (blue) dominates. This pattern suggests that younger researchers, who are likely at the beginning or mid-stages of their careers, feel more influenced by the ASN compared to their older counterparts. Similarly, the bibliometric-based chart (second figure) demonstrates that younger researchers (<35 and 35–44 age groups) also show the highest measurable productivity increases (red). However, a notable difference emerges in older age groups (45–54 and 55–65), where a higher proportion of the "ASN effect" is observed compared to the survey results. Even in the >65 group, a significant proportion of the "ASN effect" is evident in the bibliometric data.
Figure 3 presents the distribution of the ASN effect ("ASN effect") and no ASN effect ("No ASN effect") by geographical area (South, Center, and North).

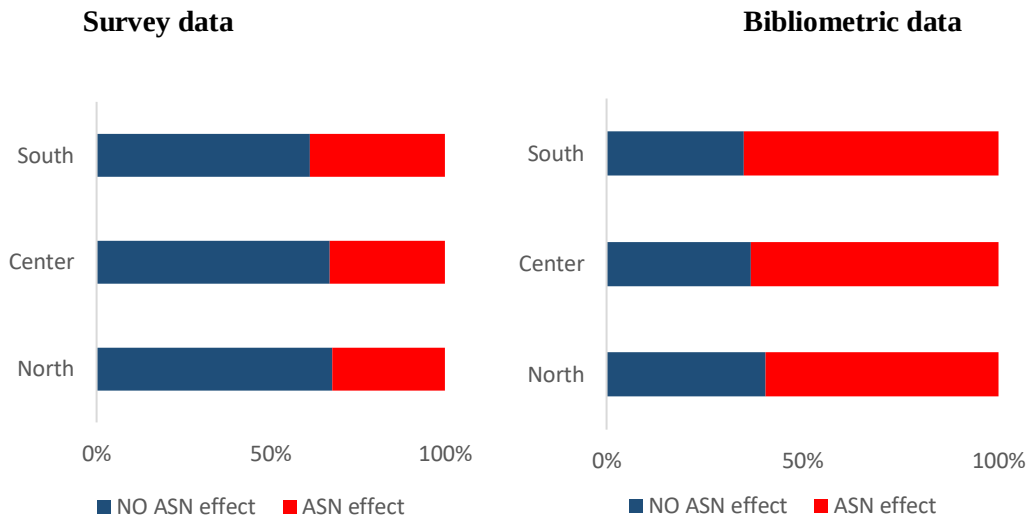


Figure 3. ASN effect by geographical area.

In the survey data, the majority of respondents indicate that the ASN has had no effect on their scientific productivity across all three regions. However, the proportion of respondents reporting an ASN effect (red) appears to be slightly higher in the South as compared to the Center and the North, suggesting that researchers in this macro-region perceive a stronger influence of the ASN on their academic output. The bibliometric data, on the other hand, present a different trend, potentially indicating a stronger measurable ASN effect across regions.
Figure 4 presents the distribution of the ASN effect and no ASN effect by university size.

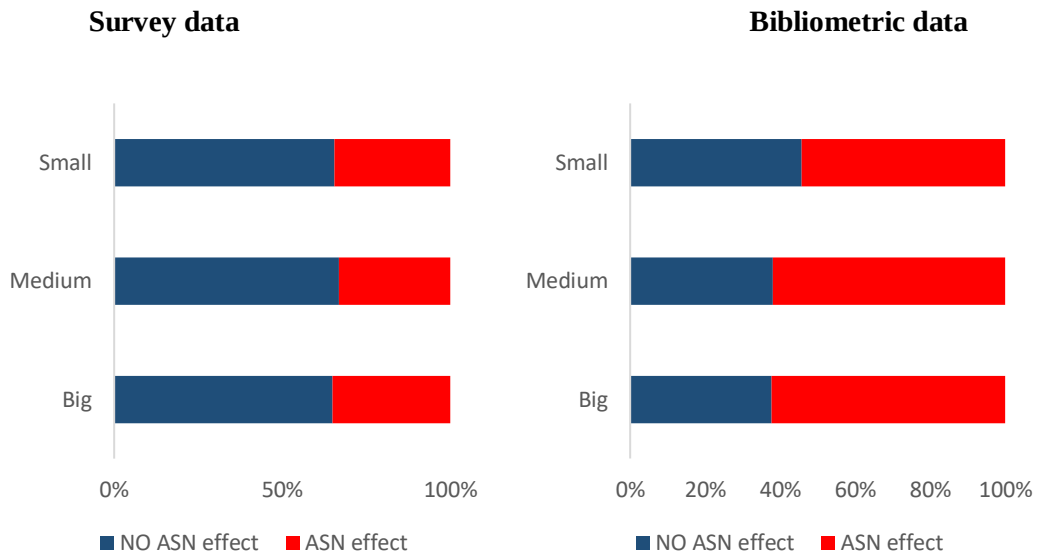


Figure 4. ASN effect by university size.

In the survey data, the majority of respondents across all university sizes (small, medium, and big) indicate that the ASN has had no effect on their scientific productivity. The bibliometric data reveals that the ASN effect appears to be more significant across all university sizes, indicating a measurable increase in scientific productivity.

Finally, Figure 5 illustrates the distribution of the ASN effect ("ASN Effect") and no ASN effect ("No ASN Effect") across different CUN categories (CUN 1 to CUN 11) for both survey-based and bibliometric data.

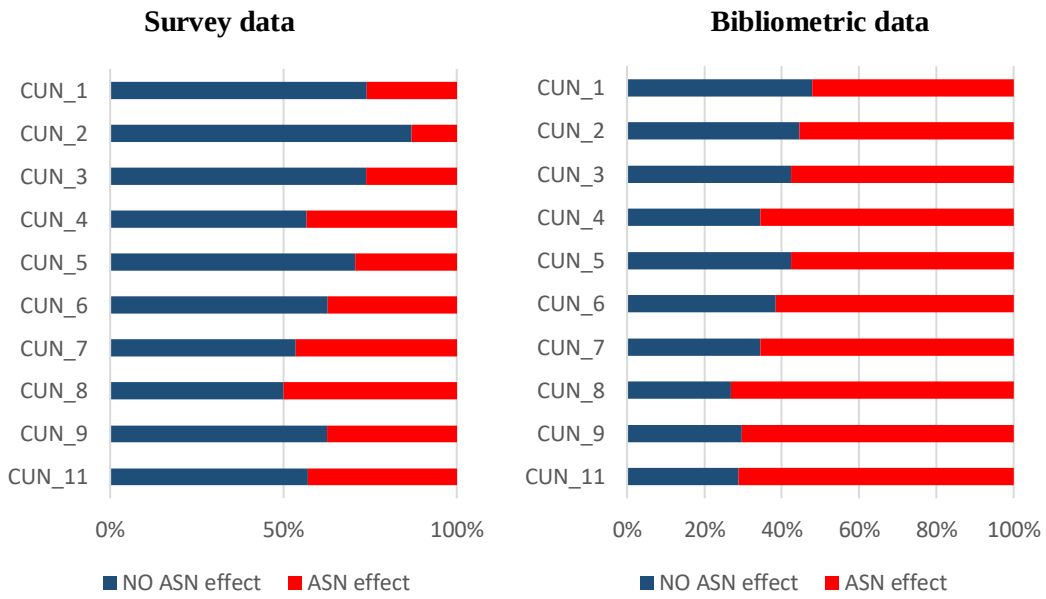


Figure 5. ASN effect by disciplinary area.

In the survey-based chart, some CUN areas, such as CUN 7 (Agricultural and Veterinary Sciences) and CUN 8 (Civil Engineering and Architecture), show a relatively high proportion of respondents reporting an "ASN Effect" (red). Conversely, fields like CUN 1 (Mathematics and Informatics), CUN 2 (Physics), and CUN 3 (Chemistry) report lower levels of perceived impact, with the "No ASN Effect" (blue) dominating. The bibliometric chart presents a different perspective on the ASN effect, reflecting changes in scientific productivity. Fields such as CUN 1 (Mathematics and Informatics) and CUN 2 (Physics) display lower proportions of the "ASN Effect" in terms of measurable increases in productivity. In contrast, CUN 8 (Civil Engineering and Architecture) shows a higher difference in productivity between the two periods (2008–2012 and 2013–2017), indicating a stronger bibliometric impact of the ASN.

The descriptive analysis highlights differences between survey-based perceptions and bibliometric evidence regarding the impact of the ASN on scientific productivity. While the majority of surveyed researchers report no significant effect of the ASN, bibliometric data suggest a measurable increase in productivity.

The econometric model

Survey

We applied a logit model, a statistical technique used to examine the relationship between a binary outcome variable and one or more predictor variables. Specifically, it models the log odds of the binary outcome as a linear function of the predictors and employs a logistic function to estimate the probability of the outcome being 1—in this case, whether the ASN influenced academics' scientific productivity. The logit model incorporates the survey's methodological design, including the sampling process. Specifically, sampling weights were included in the analysis to account for the probability of each observation being selected. These weights were also used to adjust for nonresponse and ensure that the estimates reflect the characteristics of the target population. To ensure consistency with the focus of this study, only academics who had participated in at least one ASN evaluation cycle were involved and invited to answer this question. The productivity perceived by the respondents refers to their subjective evaluation of the impact of the ASN on their scientific output, varying according to their disciplinary field. This perception includes aspects such as the volume of publications, the effort required to align with ASN standards, and the prioritization of specific research outputs.

The following table 3 presents the results of the logistic regression where "Productivity" is the dependent variable (1: ASN effect on productivity - 0: No ASN effect on productivity).

Table 3. Logistic regression on the effect of ASN (1) vs. no effect of ASN (0) on research productivity: evidence from survey data.

	Coef.	Std Err.	[95% Conf. Interval]	
Gender (1=male;0=female)	-0.330**	(0.166)	-0.655	-0.006
Age: <35	1.580**	(0.643)	0.319	2.840
Age: 35 - 44	1.820***	(0.356)	1.122	2.518
Age: 45 - 54	1.149***	(0.345)	0.473	1.826
Age: 55 - 65	1.020***	(0.341)	0.351	1.688
Univ. Medium Size vs Univ. Large Size	-0.129	(0.179)	-0.480	0.222
Univ. Small Size vs Univ. Large Size	-0.0396	(0.222)	-0.475	0.395
Geo: North vs Center	-0.100	(0.195)	-0.482	0.282
Geo: South and Islands vs Center	0.235	(0.215)	-0.187	0.657
Cun Area 1 vs Cun Area 2	0.834*	(0.478)	-0.102	1.770
Cun Area 3 vs Cun Area 2	0.801	(0.495)	-0.170	1.771
Cun Area 4 vs Cun Area 2	1.751***	(0.557)	0.658	2.843
Cun Area 5 vs Cun Area 2	1.061**	(0.440)	0.198	1.923
Cun Area 6 vs Cun Area 2	1.547***	(0.435)	0.695	2.399
Cun Area 7 vs Cun Area 2	1.798***	(0.454)	0.907	2.688
Cun Area 8 vs Cun Area 2	1.912***	(0.513)	0.906	2.918
Cun Area 9 vs Cun Area 2	1.405***	(0.434)	0.555	2.255
Cun Area 11 vs Cun Area 2	1.680***	(0.577)	0.550	2.811
Constant	-2.871***	(0.553)	-3.954	-1.788

*** p<0.01, ** p<0.05, * p<0.1

The regression results highlight several important patterns regarding the factors influencing perceptions of productivity increases attributed to the ASN. Gender plays a significant role, with male academics being less likely than their female counterparts to report that the ASN has positively impacted their productivity. Age also emerges as a crucial factor, with younger academics, particularly those under 35 and in the 35–44 age range, significantly more likely to report productivity increases due to the ASN. This suggests that early-career researchers, who are often more dependent on evaluation systems for career progression, are more responsive to the pressures and incentives created by the ASN. However, senior academics, while slightly less likely to attribute an impact compared to younger colleagues, also report the effects of the ASN on their productivity. This indicates that the influence of the ASN is not limited to any single career stage but is felt across all age groups, underscoring its pervasive impact on academic publishing behaviors.

The results show considerable variation when considering disciplinary differences (as represented by CUN areas). Academics in certain fields, such as those in Areas 4 (Earth Sciences), 5 (Biology), 6 (Medicine), 7 (Agricultural and Veterinary Sciences), 8 (Civil Engineering and Architecture), and 9 (Industrial and Information Engineering), are significantly more likely to attribute productivity increases to the ASN compared to those in Area 2 (Physics). This suggests that disciplines with

different publication practices and evaluation standards may respond differently to the incentives of the ASN, with some fields feeling a stronger push to align their outputs with its requirements.
Geographic location and university size do not show a significant effect on the likelihood of reporting productivity increases attributed to the ASN. This suggests that the likelihood of reporting productivity increases attributed to the ASN appears to be primarily influenced by individual characteristics or specific area-based indicators rather than by institutional factors.

Bibliometric analysis

The logistic regression model presented here examines the factors influencing the likelihood of observing a measurable increase in scientific productivity attributed to the ASN, as determined by bibliometric data. The dependent variable is binary, taking the value of 1 if there is a measurable increase in productivity between the two periods (2008–2012 and 2013–2017) and 0 otherwise.
Table 4 presents the results of the logistic regression, characterized by the following features.
Number of obs = 26217
Wald chi2(18) = 742.25
Prob > chi2 = 0.0000
Log pseudolikelihood = -17013.07
Pseudo R2 = 0.0222

Table 4. Logistic regression on the effect of ASN (1) vs no effect of ASN (0) on research productivity: evidence from bibliomrtric data.

	Coef.	**Std Err.**	**[95% Conf. Interval]**	
Gender (1=male;0=female)	0.015	0.029	-0.041	0.071
Age: <35	1.100***	0.288	0.535	1.665
Age: 35 - 44	0.841***	0.21	0.428	1.253
Age: 45 - 54	0.578***	0.209	0.167	0.988
Age: 55 - 65	0.253	0.209	-0.157	0.663
Univ. Medium Size vs Univ. Large Size	-0.062**	0.028	-0.116	-0.007
Univ. Small Size vs Univ. Large Size	-0.417***	0.14	-0.692	-0.142
Geo: North vs Center	-0.214***	0.032	-0.275	-0.152

Geo: South vs Center	0.026	0.035	-0.042	0.095
Cun Area 1 vs Cun Area 2	-0.871***	0.087	-1.042	-0.7
Cun Area 3 vs Cun Area 2	-0.679***	0.093	-0.86	-0.497
Cun Area 4 vs Cun Area 2	-0.647***	0.088	-0.82	-0.475
Cun Area 5 vs Cun Area 2	-0.269**	0.108	-0.48	-0.057
Cun Area 6 vs Cun Area 2	-0.591***	0.084	-0.756	-0.427
Cun Area 7 vs Cun Area 2	-0.370***	0.081	-0.529	-0.211
Cun Area 8 vs Cun Area 2	-0.297***	0.088	-0.47	-0.123
Cun Area 9 vs Cun Area 2	0.055	0.102	-0.144	0.254
Cun Area 11 vs Cun Area 2	-0.104	0.085	-0.271	0.063
Constant	0.490**	0.222	0.054	0.926

*** p<0.01, ** p<0.05, * p<0.1

Observing the results of bibliometric regression, we see that Age emerges as a strong predictor, with younger academics, particularly those under 35 and in the 35–44 age range, significantly more likely to report the outcome under consideration compared to older colleagues. While the likelihood decreases with age, academics aged 45–54 also show significant effects. However, for those aged 55–65, the effect is no longer statistically significant, suggesting that the influence of this factor diminishes with seniority.

Institutional size plays an important role, with academics affiliated with medium-sized and small universities being less likely to report the outcome compared to those at large universities. This effect is particularly pronounced for small universities, where the likelihood of reporting the outcome is significantly reduced. These findings indicate that institutional environments at larger universities may create easier conditions for achieving the specified outcome.

Geographic differences also emerge, with academics in the North of Italy being significantly less likely to report the outcome compared to those in the Center. However, no significant differences are observed between the South and Center, suggesting a more uniform experience in those regions.

The results reveal considerable variation across disciplinary areas. Academics in Areas 1 (Mathematics and Informatics), 3 (Chemistry), 4 (Earth Sciences), 5 (Biology), 6 (Medicine), 7 (Agricultural and Veterinary Sciences), and 8 (Civil

Engineering and Architecture) are significantly less likely to report the outcome compared to those in Area 2 (Physics). Notably, Area 9 (Industrial and Information Engineering) and Area 11 (only Psychology sub-area) do not show significant differences compared to Area 2, suggesting closer alignment in these disciplines.

The comparison between the survey and bibliometric analyses reveals both convergences and divergences in the factors shaping the impact of the ASN on scientific productivity in STEMM fields. Both approaches underscore the strong influence of age, with younger academics, particularly those under 35 and in the 35–44 age range, significantly more likely to report productivity increases attributed to the ASN. This suggests that early-career researchers are more responsive to the ASN's incentives for career progression. Institutional size, however, emerges as a significant factor only in the bibliometric analysis, where academics at medium-sized and small universities report fewer productivity increases compared to their counterparts at larger universities. This likely reflects disparities in resources, access to academic knowledge networks and research infrastructure. These constraints can make it harder to align with ASN-driven incentives, particularly in fields where collaboration and resource intensity are critical for publishing high-quality work.

Regarding the geographic location, the bibliometric analysis identifies lower effects in the North of Italy, while the survey finds no significant regional differences. Both analyses highlight disciplinary differences, although in contrasting directions: the survey identifies stronger effects in fields such as Earth Sciences, Biology, and Medicine, while physicists report being less influenced by the ASN, suggesting that their perceived increase in productivity is less tied to the evaluation tool. In contrast, the bibliometric analysis indicates that Physics, taken as the baseline category in the model, shows higher productivity increases compared to other disciplines. This discrepancy suggests that while physicists do not attribute their increased productivity to the ASN in the survey, the bibliometric evidence points to an actual increase in their output, which may instead be driven by other factors, such as intrinsic disciplinary dynamics and stronger collaboration networks.

Conclusions

This study highlights a significant misalignment between researchers' perceptions of productivity increases attributed to the Italian Scientific Habilitation (ASN) and the evidence obtained through bibliometric assessments. While the bibliometric analysis reveals that the majority of academics (62%) experienced measurable increases in scientific productivity following the introduction of the ASN, survey data indicate that most researchers (65%) perceive little to no effect on their productivity. This discrepancy underscores a fundamental difference in how the effects of the evaluation systems are experienced versus their quantifiable outcomes.

The divergence between perception and evidence is particularly notable across demographic and contextual factors. Younger researchers and those at earlier stages of their careers are more likely to report productivity increases, both in survey responses and bibliometric data, reflecting their stronger dependence on evaluation systems for career progression. However, in older age groups, while bibliometric evidence points to measurable productivity increases, these are often not recognized

or attributed to the ASN by the researchers themselves. Similarly, disciplinary differences reveal contrasting patterns: researchers in fields such as Medicine and Engineering report and exhibit higher productivity increases, while those in Physics and Mathematics show a significant bibliometric impact but perceive less influence from the ASN.

These differences have critical implications for researchers, institutions, and policymakers. As it regards researchers, the key question is: what lies at the root of the mismatch between perception and evidence? If the discrepancy stems from researchers failing the habilitation exercises, it could significantly undermine trust in evaluation systems, particularly among those who feel their contributions are undervalued or overlooked. In such cases, researchers might abandon efforts toward continuous improvement or resort to counterproductive behaviors. These could include prioritizing short-term outputs over long-term discoveries or interdisciplinary work, or engaging in unethical practices like excessive self-citation, citation networks, salami-slicing publications, or searching for honorary authorship.

As for institutions, universities must navigate the varying impacts of evaluation systems across disciplines and demographics. The observed disparities may suggest that a one-size-fits-all approach to research assessment is insufficient. Institutions should aim to foster environments where diverse academic contributions, including teaching, mentorship, and technology transfer, are valued alongside publications.

Talking about policymakers, the findings emphasize the need for more nuanced and inclusive evaluation policies. Efforts to improve the transparency and communication of evaluation criteria and results could help bridge the gap between perception and evidence, enhancing the legitimacy of these systems and forging researchers' virtuous behavior.

Policy recommendations stemming from this study include but are not limited to i) tailoring discipline-specific metrics that align with the unique publication practices and priorities of each field; ii) promoting transparency by clearly communicating how metrics are used in the evaluation and providing feedback to researchers on how their work aligns with institutional and national goals; and iii) addressing the equity gaps by implementing targeted measures to reduce disparities observed in gender, geographic location, and institutional size, ensuring fair and equitable evaluation processes.

Finally, this research underscores the need for ongoing dialogue among policymakers, institutional leaders, and researchers to ensure that evaluation systems align with academic values and societal goals. By bridging the gap between perception and evidence, we can foster trust, inclusivity, and innovation in the academic community, ensuring that research evaluation serves its ultimate purpose: advancing knowledge and addressing global challenges.

Despite its contributions, this study has methodological limitations. The survey data relies on self-reported perceptions, which may be influenced by personal biases or an incomplete understanding of the factors driving productivity changes. Conversely, bibliometric analyses rely on proxies for productivity, which may overlook qualitative aspects of academic work. Furthermore, bibliometric analyses infer causality based on observed trends, which may not fully capture the complex

interplay of motivations and constraints affecting researchers. The limitations of comparing the possible mismatch between perceptions and bibliometric evidence are twofold. On the one hand, the strength of the causal attribution of changes in research productivity to the ASN is not the same. It derives from individual appreciation in the case of the survey, while it is inferred in the case of the bibliometric analysis by observing the levels and characteristics of productivity in the different fields before and after the introduction of the ASN. On the other hand, the survey also collects the perceptions of the respondents on the importance of other factors beyond the ASN on the changes in research productivity. Therefore, the attribution of the effect observed to the ASN can be calibrated with respect to other causes that played a role in the production of the effect.

Future research should explore the underlying reasons for these misalignments, incorporating mixed methods and longitudinal designs to better understand the evolving relationship between perception, evidence, and the broader academic environment.

In conclusion, addressing the gap between researchers' perceptions and bibliometric evidence is essential for building trust and ensuring that evaluation systems serve their intended purpose. By aligning these systems with academic values and promoting inclusivity, we can foster environments that support both individual and collective advancement in knowledge creation.

References

Abramo, G. (2024). The forced battle between peer-review and scientometric research assessment: Why the CoARA initiative is unsound. *Research Evaluation*, rvae021, doi.org/10.1093/reseval/rvae021.

Abramo, G., & D'Angelo, C.A. (2021). The different responses of Italian universities to introduction of performance-based research funding. *Research Evaluation*, 30(4), 514–528.

Abramo, G., & D'Angelo, C.A. (2014). How do you define and measure research productivity? *Scientometrics*, 101(2), 1129-1144.

Abramo, G., D'Angelo, C.A., & Rosati, F. (2013). Measuring institutional research productivity for the life sciences: the importance of accounting for the order of authors in the byline. *Scientometrics*, 97(3), 779-795.

Archambault, É., Vignola-Gagné, É., Côté, G. et al. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics* 68, 329–342.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45–80.

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, 7, 291.

de Diego, I.M., Prieto, J.C., Fernández-Isabel, A., Gomez, J. & Alfaro, C. (2024). Framework for scoring the scientific reputation of researchers. *Knowledge and Information Systems*, 66, 3523–3545.

de Rijcke, S., Wouters, P.F., Rushforth, A.D., Franssen, T.P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation*, 25(2), 161–169.

DORA. (2012). Declaration on Research Assessment.

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. *GigaScience*, 8(6), giz053.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471–479.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429–431.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572.

Ioannidis J.P.A., & Maniadis, Z. (2023). In defense of quantitative metrics in researcher assessments. *PLoS Biology*, 21(12): e3002408.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature News*, 504(7479), 211.

McKiernan, E. C., et al. (2016). How open science helps researchers succeed. *eLife*, 5, e16800.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer Science & Business Media.

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P., & Goodman, S. N. (2022). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, 20(3), e3001606.

Mongeon, P., Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106, 213–228.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Nicholas, D., Watkinson, A., Boukacem-Zeghmouri, C., Rodriguez-Bravo, B., Xu, J., Abrizah, A., & Herman, E. (2017). Early career researchers and their publishing and authorship practices. *Learned Publishing*, 30(2), 101–112.

Sahel, J. A. (2011). Quality versus quantity: Assessing individual research performance. *Science Translational Medicine*, 3(84), 84cm13.

Seeber, M., Cattaneo, M., Meoli, M., & Malighetti, P. (2019). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, 48(2), 478-491.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), 160384.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2), 365–391.

https://doi.org/10.51408/issi2025_044

The Impact of Russia-Ukraine Conflict on International Migration of Russian-Affiliated Researchers

Andrey Lovakov

lovakov@dzhw.eu
German Centre for Higher Education Research and Science Studies (DZHW), Schützenstraße 6A, 10117 Berlin (Germany)

Abstract

The Russia-Ukraine conflict has had a significant impact on international migration patterns, including a significant exodus of Russian-affiliated researchers. This study examines the scale, disciplinary impact, and geographic shifts of this migration wave by analyzing data from the Scopus database. Using changes in the most frequent country of affiliation as a proxy for migration, the results show a substantial decline in the net migration rate of Russian researchers from 2022 to 2024. Russia has been losing about 0.8% of its active researchers annually over this period. This brain drain wave affects almost all research fields. The most affected disciplines include Physics and Astronomy, Computer Science, and Mathematics, while Dentistry and Health Professions experienced comparatively smaller declines. Geographically, traditional academic destinations such as Germany, the United States, and Switzerland have absorbed the majority of emigrating researchers, while non-traditional destinations, such as Armenia, the United Arab Emirates, and Kazakhstan, are also becoming important. However, large academic systems such as China and India have not seen significant increases. The findings underscore that this unprecedented brain drain will have both short- and long-term consequences for Russian academia and global science.

Introduction

The Russia-Ukraine conflict has dramatically reshaped the geopolitical, economic, and social landscape, with significant implications for international migration patterns. While migration from Ukraine has mainly taken the form of refugee movements in search of immediate safety, migration from Russia has different drivers. Economic sanctions, growing political repression, fear of conscription following Russia's mobilization campaigns, and moral opposition to the conflict have led many Russian citizens to flee abroad. For the academic and research community, these factors are compounded by concerns about academic freedom, the sustainability of international collaborations, and the narrowing space for intellectual dissent.

While data on the exact scale of researcher migration from Russia remains scarce, emerging evidence suggests a broader trend of intellectual flight. Wachs (2023), for example, documented a notable shift among Russian open-source software developers: 11.1% listed a new country on GitHub by November 2022, compared to only 2.8% of developers from neighboring countries not involved in the conflict. Similar trends are likely to exist within other segments of Russia's intellectual community, including academic researchers. Chankseliani and Belkina (2024) provided an overview of various estimates of the outflow of researchers from Russia. For example, an analysis based on the ORCID database estimated that about 2500 scientists left Russia after February 2022, when the armed conflict began. However,

it is important to note that the ORCID database relies on self-reported information and has limited coverage of the Russian research community.

The current study aims to examine the impact of the Russia-Ukraine conflict on the international mobility of Russian-affiliated researchers, focusing on the scale of migration, the disciplines most affected, and the primary destinations of these migrating researchers. Understanding the extent and characteristics of this migration is important for several reasons. First, it sheds light on the broader consequences of the conflict for global scientific networks, particularly in fields where Russian researchers have traditionally been active contributors (such as physics, mathematics, chemistry (Lovakov, 2022)). Second, it provides valuable information to receiving countries, which may consider adopting targeted policies to attract and support displaced researchers. Third, it contributes to a deeper understanding of how geopolitical crises influence the mobility of intellectual communities, with implications for both policy and practice. Chankseliani and Belkina (2024) noted that this wave of migration may be different from the previous one that followed the dissolution of the Soviet Union. While the previous wave of migration was influenced by an evolving political or economic landscape and economic drivers (Ganguli, 2014; Graham & Dezhina, 2008; Yegorov, 2009), the current wave is more immediate and driven not only by economic reasons, but also by personal safety concerns and opposition to government actions. This specificity requires a better understanding of the current wave of migration and its potential impact on the Russian and global academic system.

This study seeks to answer the following research questions: 1) How many researchers have left Russia in response to the armed conflict? 2) Which academic disciplines have been most affected by this migration? 3) What are the main destination countries for Russian researchers?

Method

Data

The in-house version of the Scopus database provided by the German Competence Network for Bibliometrics (snapshot as of 01.2025) was used (Schmidt et al., 2024). All authors who affiliated with Russian institution in at least one publication indexed in Scopus were selected. There are 856,853 author profiles of researchers who have published at least once with a Russian address in the period 1996-2024. For each of these authors all its publications and affiliations were found. There are 3,575,868 publications in Scopus published between 1996 and 2024 associated with these authors. The Scopus author ID was used to identify all publications for each author. All affiliations and publications associated with the same author ID were considered to be affiliations and publications of the one same author. It was shown that Scopus data and Scopus author ID are suitable to identify the international mobility of a scientist and could be a good solution (Aman, 2018; Baas et al., 2020).

Migration event

The data include the year of publication, the address, and the country, that can be used as a proxy for the author's residential addresses. To detect a migration event,

the most frequent (mode) country of affiliation is extracted for each researcher in each year. A migration event is considered to have occurred if the researcher's most frequent country of affiliation changes in two different years. This so-called "mode-based method" is a widely used method for identifying migration events (Akbaritabar et al., 2024; Subbotin & Aref, 2021; Zhao et al., 2022). When there were two the most frequent countries, they were compared to the most frequent country in the previous year. If one of them is the same as the previous year, that country was selected as the country of residence. If none of them matched the most frequent country in the previous year, one of them was chosen at random. When available, the year of "early access", "online first" or "in press" was used to more precisely identify the time of the migration event.

Research field assignment

Each author has been classified into one of the Subject Area Classifications, which are based on the All Science Journal Classification (ASJC) scheme. Each serial title in Scopus is classified into one or more subcategories of the ASJC. The 334 lower-level subcategories are assigned to one of the 27 top-level fields. Each individual publication can also be assigned to one of these fields. To assign an author to a research field, the most frequent (mode) research field was extracted. If there were two or more most frequent fields, one of them was chosen at random.

Measures

Based on these data, several measures were calculated. In-migration I_y was calculated as the number of published researchers who immigrated to Russia in year y. Out-migration E_y was calculated as the number of published researchers who emigrated from Russia in year y. The estimated population of researchers in Russia in year y (M_y) was calculated as the number of researchers with Russia as the mode country of affiliation. If an author does not publish every year, we assume that he or she is still part of the population of active researchers two years before the nearest publication year. Only authors with a total of more than one Scopus-indexed publication in their entire career were included in these calculations. Authors with only one Scopus-indexed publication were not considered as active members of the academic community. Net migration rate NMR_y was calculated as the difference between in-migration and out-migration rates per 100 researchers:

$$NMR_y = (I_y - E_y)*100/M_y.$$

The main idea of the analysis is to compare the net migration rate in the last three years (2022-2024) with the net migration rate in previous years.

Results

Figure 1 shows the inflow, outflow, and net migration rate per 100 researchers in Russia from 1997 to 2024. Overall, the net migration rate remained negative for most of the study period, indicating a persistent net outflow of researchers. From 1997 to 2014, the net migration rate showed a gradual improvement, starting at -0.36 in 1997 and peaking at -0.05 in 2014. This trend coincides with an increase in the number of

active researchers (from 74,829 in 1997 to 141,035 in 2014). A turning point occurred in 2015, when the net migration rate became positive for the first time (0.12). However, in 2018, the net migration rate returned to negative values, indicating a resurgence of net emigration, which will gradually worsen until 2023. The period from 2022 to 2024 shows a steep decline in the net migration rate, falling from -0.17 in 2021 to -0.83 in 2023, the second lowest value in the dataset. This dramatic drop coincides with the onset of the armed conflict between Russia and Ukraine, which is likely to exacerbate emigration (1,465 and 1,938 researchers emigrated in 2022 and 2023, compared to only 672 and 618 who immigrated). It means that Russia has been losing about 0.8% active researchers per year for the last two years. In addition, the shrinking population of active researchers also points to broader structural problems in the academia. It is possible that researchers have either left academia or are still in the process of looking for an academic position. It is important to note, however, that the coverage of 2024 may is not complete and that all measures may change in the future when all publications are included.

Figure 2 shows the net migration rates in different research fields. The period 2022-2024 shows a significant decrease in the net migration rate in almost all research fields in Russia. The largest net migration rate are observed in Physics and Astronomy (-4.42 per 100 researchers), Computer Science (-3.45), Economics, Econometrics and Finance (-2.89), Business, Management and Accounting (-2.62), Mathematics (-2.38), Neuroscience (-2.35), Social Sciences (-2.06), Biochemistry, Genetics and Molecular Biology (-1.83), Psychology (-1.56), and Arts and Humanities (-1.53). Whereas the lowest net migration rate is observed in Medicine (-0.34). Overall, the data show that high-technology and internationally integrated disciplines, such as Physics, Mathematics, and Computer Science, are the most affected by brain drain.

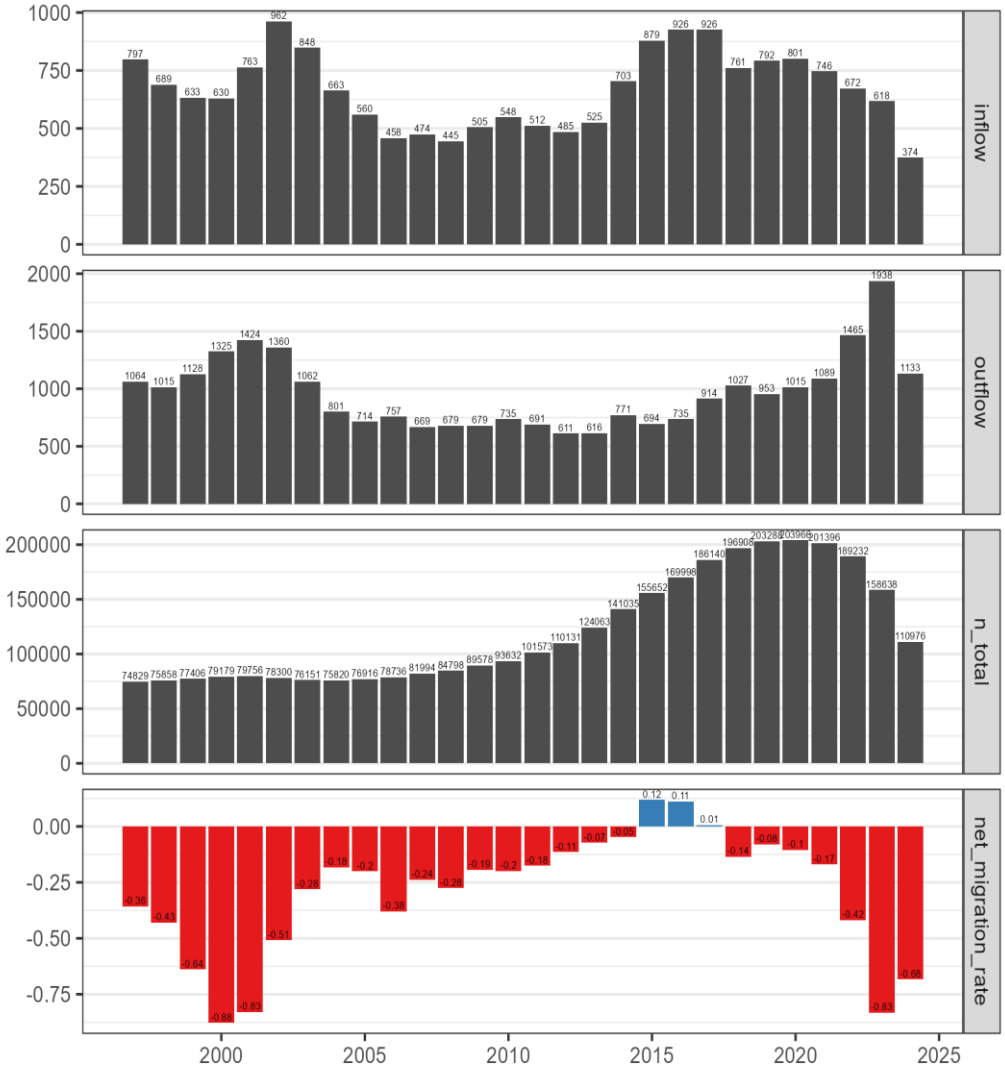


Figure 1. In-migration, out-migration, and net migration rate per 100 researchers in Russia over the 1997–2024 period.

An analysis of the net migration rate by countries reveals some significant changes for some countries (see Figure 3). Compared to earlier periods, the net migration rate shows a marked increase in out-migration from Russia for several destination countries. Traditional destinations for Russian researchers such as Germany, the United States, Switzerland, Finland, Israel, experienced the most notable increase in last three years. Switzerland's net migration rate dropped from -0.018 in 2019–2021 to a dramatic -0.294 in 2022–2024. This drop is most likely due to CERN's policy regarding Russian affiliated researchers. Europe's particle-physics laboratory CERN has decided not to renew agreements with Russia and Belarus when they expire in 2024. All Russian-affiliated scientists should have lost access to the CERN site and must hand in any French or Swiss residency permits they hold after November 2024

(Gibney, 2024). All Russia-affiliated scientists who wanted to continue working on CERN's projects had to find positions in institutions outside of Russia. Our results show that many researchers apparently succeeded in doing so.

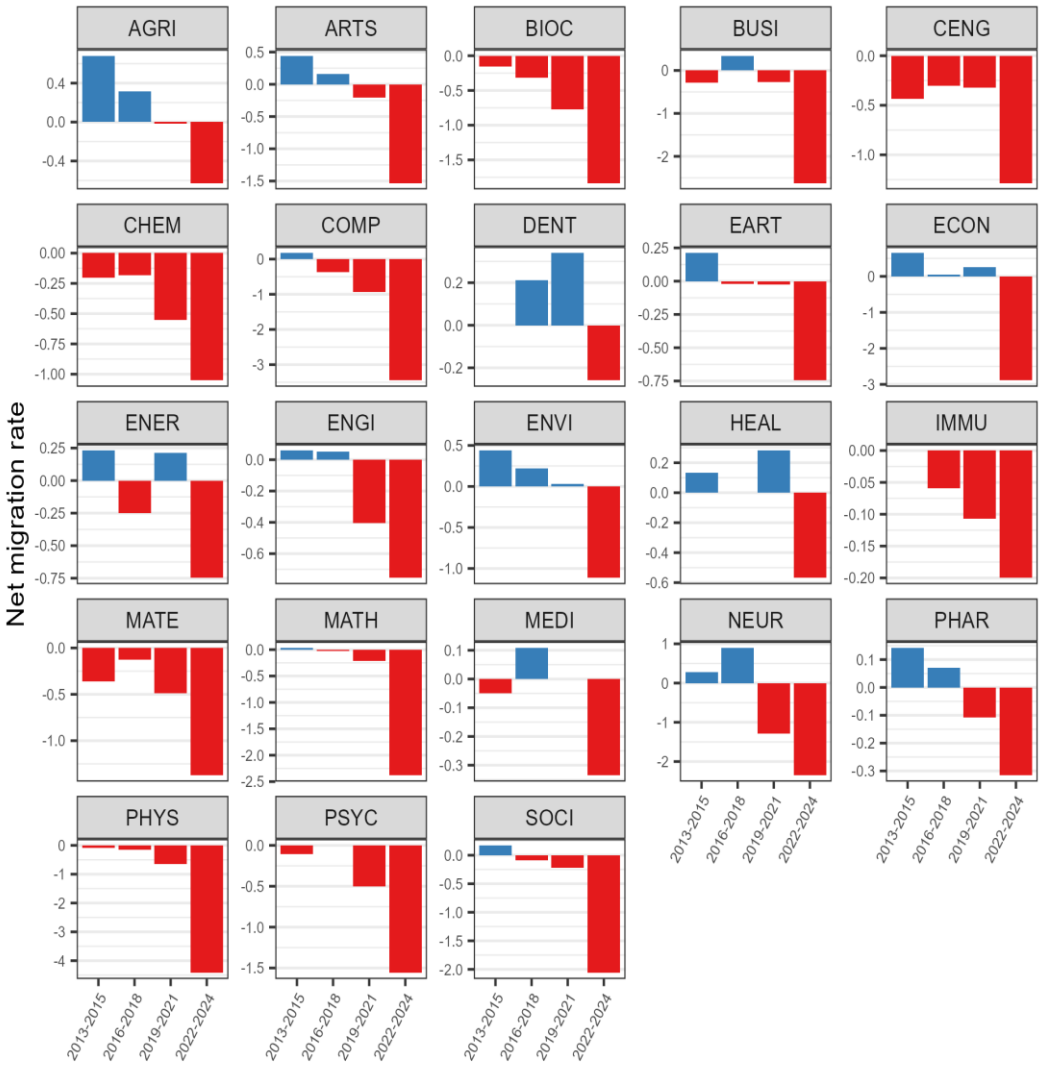


Figure 2. Net migration rate per 100 researchers in Russia in different fields. Nursing, Veterinary, and Decision Sciences are not shown because of small total number of authors (< 100 authors).

Similarly, the United Kingdom, France, and other European countries also increased the number of immigrated Russian researchers. Italy rose from 33rd to 10th place in the ranking of countries with the highest outflow from Russia, and Spain moved from 27th to 12th place, Canada moved from 32nd to 13th place. These changes suggest a diversification of emigration destinations, possibly due to the availability of academic opportunities and a welcoming environment for skilled migrants in these countries.

Other significant risers include countries that were not typical destinations for academic mobility. Armenia rose from 31st to 16th place, Kazakhstan rose from 18th to 6th place, the United Arab Emirates rose from 28th to 17th place. Interestingly, China and India, which are among the largest academic systems in the world, show relatively smaller changes. Russian researchers are not actively moving to these countries. The reasons for this should be investigated in the future. Overall, these results show that Russian researchers are mainly moving to typical destinations in Europe and the US, however there are also some new destinations that were not very attractive in previous periods.

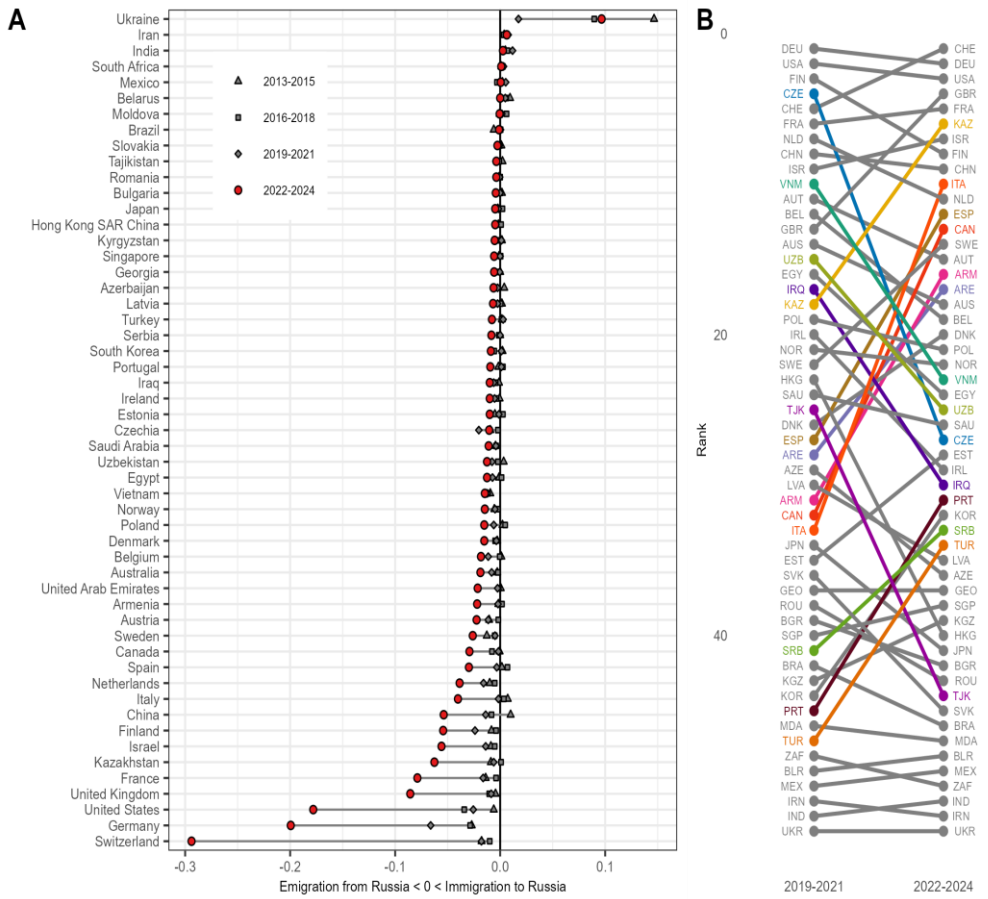


Figure 3. A - Changes in Russia's net migration rate per 100 researchers with different countries (only countries with more than 30 outcome or income researchers in sum between 2013-2024 are shown). B – Ranking of countries based on the net migration rate per 100 researchers with Russia (the higher the rank, the higher the emigration from Russia).

Discussion and Conclusion

The results of this study provide important insights into the impact of the Russia-Ukraine conflict on the international migration patterns of Russian-affiliated researchers. The sharp decline in the net migration rate in 2022-2024 highlights the

magnitude of this phenomenon, with Russia losing about 0.8% of its active researchers per year during this period. There is both an increase in emigration and a decrease in immigration. However, there is also a huge decrease in the total number of active researchers in Russia over the last three years. This huge decrease could also be a sign of future academic emigration. Some researchers may still be in the process of looking for a new position, or they may have moved but haven't published yet because it takes time to start a new project, prepare and publish new papers based on that new project. This unprecedented loss of researchers has far-reaching implications, not only for the Russian academic system, but also for global scientific networks, host countries, and the disciplines most affected by this migration.

The results underline that this brain drain wave affects almost all research fields. However, the most internationally integrated and high-tech disciplines, such as Physics, Mathematics, and Computer Science, are more severely affected. These fields have historically been stronger and more internationally oriented in Russia, with greater opportunities for academic mobility and emigration.

The study also highlights shifts in the geography of academic migration from Russia. Traditional destinations for Russian researchers, such as Germany, the United States, and Switzerland, continue to attract significant numbers of emigrants, confirming their status as hubs of global academic mobility. However, the rise of non-traditional destinations such as Armenia, the United Arab Emirates, and Kazakhstan signals a diversification of migration patterns. This diversification may reflect a combination of factors, including the geopolitical landscape, visa and migration policies, and the availability of academic opportunities in these countries. For host countries, this trend offers opportunities to strengthen their research ecosystems by attracting highly qualified talent. It also highlights the importance of creating a supportive environment for displaced researchers, including funding and integration programs. Interestingly, large academic systems such as China and India have not yet experienced a significant influx of Russian researchers. This could be due to both linguistic and cultural barriers, limited compatibility between academic systems, or political factors. Future research should explore the underlying factors that make certain destinations more attractive or less attractive to migrating researchers, especially in the context of global and regional geopolitical dynamics.

For Russia, the findings reveal a deepening crisis within its academic system. The declining number of active researchers, coupled with a significant brain drain, is weakening the academic system and creating major challenges for higher education. The immediate loss of a huge amount of talent leads to a decline in research capacity, innovation potential, and global academic standing (Chankseliani & Belkina, 2024). Addressing these issues will be a long-term challenge.

The Russia-Ukraine conflict has triggered a significant outflow of Russian-affiliated researchers, reshaping global patterns of academic mobility. This study contributes to a broader understanding of how geopolitical crises shape intellectual mobility in today's highly internationalized and mobile academic system. The current wave of migration, driven largely by political reasons and security concerns, differs from previous patterns. Its impact on global scientific networks, particularly in terms of disrupted collaborations and shifts in research priorities, requires further study.

Acknowledgments

Data for Scopus in this study were obtained from the German Competence Network for Bibliometrics (https://bibliometrie.info/), funded by the German Federal Ministry for Education and Research (BMBF) with grant number 16WIK2101A.

References

Akbaritabar, A., Theile, T., & Zagheni, E. (2024). Bilateral flows and rates of international migration of scholars for 210 countries for the period 1998-2020. *Scientific Data, 11*(1), 816. https://doi.org/10.1038/s41597-024-03655-9

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics, 117*(2), 705–720. https://doi.org/10.1007/s11192-018-2895-3

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies, 1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Chankseliani, M., & Belkina, E. (2024). Academic Exodus from Russia: Unravelling the Crisis. *Journal of Comparative & International Higher Education, 16*(3), Article 3. https://doi.org/10.32674/jcihe.v16i3.6304

Ganguli, I. (2014). Scientific Brain Drain and Human Capital Formation After the End of the Soviet Union. *International Migration, 52*(5), 95–110. https://doi.org/10.1111/imig.12165

Gibney, E. (2024). CERN prepares to expel Russian scientists—But won't completely cut ties. *Nature*. https://doi.org/10.1038/d41586-024-02982-6

Graham, L. R., & Dezhina, I. (2008). *Science in the New Russia: Crisis, Aid, Reform*. Indiana University Press.

Lovakov, A. V. (2022). Disciplinary Structure of Scientific Research in the Post-Soviet Countries. *Automatic Documentation and Mathematical Linguistics, 56*(6), 275–284. https://doi.org/10.3103/S000510552206005X

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., Taubert, N., Bausenwein, T., & Stahlschmidt, S. (2024). *The Data Infrastructure of the German Kompetenznetzwerk Bibliometrie: An Enabling Intermediary between Raw Data and Analysis*. Zenodo. https://doi.org/10.5281/zenodo.13935407

Subbotin, A., & Aref, S. (2021). Brain drain and brain gain in Russia: Analyzing international migration of researchers by discipline using Scopus bibliometric data 1996–2020. *Scientometrics, 126*(9), 7875–7900. https://doi.org/10.1007/s11192-021-04091-x

Wachs, J. (2023). Digital traces of brain drain: Developers during the Russian invasion of Ukraine. *EPJ Data Science, 12*(1), 14. https://doi.org/10.1140/epjds/s13688-023-00389-3

Yegorov, I. (2009). Post-Soviet science: Difficulties in the transformation of the R&D systems in Russia and Ukraine. *Research Policy, 38*(4), 600–609. https://doi.org/10.1016/j.respol.2009.01.010

Zhao, X., Aref, S., Zagheni, E., & Stecklov, G. (2022). Return migration of German-affiliated researchers: Analyzing departure and return by gender, cohort, and discipline using Scopus bibliometric data 1996–2020. *Scientometrics, 127*(12), 7707–7729. https://doi.org/10.1007/s11192-022-04351-4

https://doi.org/10.51408/issi2025_020

The Increasing Fragmentation of Global Science Limits the Diffusion of Ideas

Alexander J. Gates[1], Indraneel Mane[2], Jianjian Gao[3]

[1]*agates@virginia.ed at*
School of Data Science, University of Virginia, Charlottesville, Virginia (USA)

[2]*maneindraneel@gmail.com at*
Network Science Institute, Northeastern University, Boston, Massachusett (USA)

[3]*psp2nq@virginia.edu at*
School of Data Science, University of Virginia, Charlottesville, Virginia (USA)

Abstract

The global scientific landscape emerges from a complex interplay of collaboration and competition, where nations vie for dominance while simultaneously fostering the diffusion of knowledge on a global scale. This raises crucial questions: What underlying patterns govern international scientific recognition and influence? How does this structure impact knowledge dissemination? Traditional models view the global scientific ecosystem through a core-periphery lens, with Western nations dominating knowledge production. Here, we investigate the dynamics of international scientific recognition through the lens of citation preferences, introducing a novel signed measure to characterize national citation preferences and enabling a network analysis of international scientific recognition. We find that scientific recognition is related to cultural and political factors in addition to economic strength and scientific quality. Our analysis challenges the conventional core-periphery narrative, uncovering instead several communities of international knowledge production that are rapidly fragmenting the scientific recognition ecosystem. Moreover, we provide a comprehensive statistical model that shows this network significantly constrains the diffusion of ideas across international borders. The resulting network framework for global scientific recognition sheds light on the barriers and opportunities for collaboration, innovation, and the equitable recognition of scientific advancements, with significant consequences for policymakers seeking to foster inclusive and impactful international scientific endeavours.

Introduction

The global scientific research ecosystem is shaped by the emergent interplay between international collaboration, competition, and recognition, which collectively drive the diffusion of ideas and the cross-border flow of knowledge (Hagstrom, 1974; Chinchilla-Rodr´ıguez et al., 2019; Marginson, 2022a). Strong national research infrastructures empower nations to vie for competitive advantages in technology, economics, security, and health. Concurrently, scientific knowledge flows on a global scale, with scientific ideas disseminating from their nation of origin and influencing research around the world. This diffusion and adoption of scientific information transcends national boundaries, forming a global network of scientific recognition and influence. However, the strength of influence is not uniform across all communities, leading to status stratification where nations are differentially recognized for their scientific contributions (Moravcsik, 1985; Schott, 1998; Galvez

et al.′, 2000; Tickner, 2013; Collyer, 2014; Gomez et al., 2022). This raises two central questions to be explored: What structural patterns underlie international scientific recognition and influence? and What are the consequences of that structure for knowledge dissemination?

The prevailing theories for the structure and consequences of global scientific recognition closely mirror economic models, with a clear hierarchy and power dynamics between the "core" of scientific knowledge production and its "periphery" such that certain regions or countries dominate the production and dissemination of scientific research while others occupy a peripheral or marginalized position (Prebisch, 1962; Shils, 1975; May, 1997; King, 2004; Zelnio, 2012). This core-periphery structure is hypothesized to have important consequences for international science by hindering diverse perspectives and knowledge diffusion. The core-periphery model tends to oversimplify the complex relationships between nations, reducing influence dynamics to a binary classification of 'core' or 'periphery', while overlooking the nuances and inter-dependencies that shape global science (Schott, 1988a). By relying on this model, policy and funding decisions risk becoming skewed in favor of established centers, reinforcing existing national disparities. Core countries dominate research agendas and attract greater resources, while peripheral regions struggle to keep pace, further entrenching their marginal position in the global scientific network (Sumathipala et al., 2004; Kozlowski et al., 2022; Abramo et al., 2020; Heimeriks and Boschma, 2014).

Quantitative support for the core-periphery structure of global scientific recognition is evident across various dimensions of academic activity, including international collaboration, researcher mobility, and citation patterns. For example, international collaboration networks show that core countries have higher degrees of centrality and connectivity than periphery countries, indicating their dominant role in global science (Leydesdorff and Wagner, 2008; Zelnio, 2012; Gui et al., 2019; Choi, 2012; Wagner et al., 2015), and the global embeddedness of a nation, quantified by proportion of internationally co-authored publications, is a significant predictor of traditional scientific impact (Wagner and Jonkers, 2017). Additional analysis utilizing hierarchical clustering and dominant flow methodologies on international collaboration networks suggest that the global scientific community consists of four tiers: core, strong semi-periphery, semiperiphery, and periphery (Gui et al., 2019). Under this model, the United States consistently occupies the core, maintaining collaborations with nearly every major scientific nation, while emerging powers like China and South Korea have only recently ascended to the core. Mobility patterns also reveal that core countries attract more foreign scientists and researchers than periphery countries, suggesting their greater availability of resources and opportunities (Freeman, 2010; Scott, 2015; Adams, 1998; Urbinati et al., 2021; Bauder et al., 2018). Scott (2015) refers to this phenomenon as "hegemonic internationalisation" where internationalization becomes an extension of global inequality and the struggle for dominance, driven by competition, rankings, and the concentration of academic power in certain geopolitical centers. Analysis of raw citation networks further demonstrate that core countries generate more citations

than periphery countries, implying their higher impact and influence on scientific research (Schott, 1988b, 1998; Choi, 2012; Gomez et al., 2022). Notably, Gomez et al. (2022) draws on the existing classification of countries into core and periphery to reveal a growing disparity between the number of citations a country receives and the textual similarity of the publications they produce.

Yet, it is often argued that the core-periphery model is entrenched in a Western-centric perspective that prioritizes resources and personnel, and thus overlooks the diverse cultural influences and research priorities shaping global scientific recognition and influence (Schott, 1988b; Seth, 2009; Marginson, 2022b). As early as 1988, Schott (1988b) suggested that the coreperiphery structure is primarily attributed to the volume of a nations' scientific output which obfuscates the importance of other key factors related to ties between countries, such as geopolitical relationships, linguistic similarities, colleagueship, scientific cooperation, and educational connections. Indeed, publication output remains heavily concentrated in the United States and a few European nations, implying that most quantitative indicators of scientific recognition—such as those based on raw publication, collaboration, and citation counts—tend to be notoriously Western-centric (May, 1997; King, 2004; Gomez et al., 2022). These metrics often overlook contributions from regions with smaller output, failing to recognize the diverse intellectual contributions and local innovations that may not fit neatly within dominant Western frameworks (Anderson, 2018). These limitations highlight the need for more nuanced approaches that account for regional and contextual variations in scientific production and influence.

Recent observations challenge the Western-centric narrative, indicating that emerging scientific nations are reshaping the global landscape of scientific recognition. Countries like China, Singapore, and South Korea are increasingly disrupting the traditional dominance of Western nations, signaling a shift in the concentration of global scientific influence (Lariviere et al., 2018; Basu et al., 2018; Leydesdorff et al., 2013; Gui et al., 2019; Choi, 2012). However, there is a growing tension between two perspectives: one that focuses on individual nations' transitions from the periphery to the core, and another that critiques the vertical stratification and lower visibility of researchers from regions like Latin America, the Middle East, and East Asia. The latter perspective is best articulated by Marginson (Marginson, 2022b) who discusses "the collapse of the centre-periphery model" which he attributes to internal collaboration and regional alliances rather than through traditional engagement with Euro-American scientific hubs (Marginson and Xu, 2023). AdamsAdams (2012) further characterizes such regional collaboration as a form of mutual recognition among partners within the region, fostering the development of emerging research economies.

Despite these qualitative insights, the tension remains unresolved due to a lack of robust quantitative evidence comparing the rise of individual countries within the existing core-periphery hierarchy with the creation of distinct regional scientific communities. Quantitative analyses are crucial for determining whether these regional networks are merely reinforcing the global hierarchy or truly reshaping it.

Without data-driven comparisons, it remains unclear whether the traditional core-periphery model still applies or if a more nuanced framework is needed to capture the evolving dynamics of global scientific influence.

Here, to map the structure underlying global scientific recognition and evaluate its implications for scientific influence, we analyzed the evolution of citation networks constructed from scientific publications and geolocated by their authors' affiliations. Although citations are only one–among many–means to acknowledge scientific recognition, the accessibility and quantity of such data provide a useful perspective of how scientific influence accumulates. Specifically, our data is built from 57,558,268 papers contained in the OpenAlex publication database from 1990 to 2022, which can be attributed to the countries from which their authors are affiliated, in total capturing the output of 223 countries and independent states. We must acknowledge that the OpenAlex database has known limitations, including incomplete affiliation coverage (Zhang et al., 2024) and a primary focus on English-language journals, which may introduce a selection bias towards Western countries (Gong et al., 2019). Despite these constraints, our results effectively identify significant patterns in scientific recognition. We then extracted 242 million citation relationships and calculated the number of country-specific citations to each paper within 5 years of publication (see SI, section S2).

To quantitatively capture scientific recognition, we adopt a popular measure of rank overrepresentation or under-representation (Methods, and SI, section S3), which empowers us to measure when one nationality over- or under-cites the papers from another nationality, accompanied by a level of statistical significance. To our knowledge, this marks the first application of such a method to determine whether one nation exhibits a preference or aversion towards another's scientific publications. The recognition relationship between countries in scientific output is influenced by a complex interplay of factors, including nationality bias, disparities in research quality, and international collaboration. Our study, through this measure, aims not to disentangle these individual factors but to elucidate the overall landscape resulting from their combined effects. We compare these citation patterns to a baseline constructed from the citation distribution of the source country to all other countries in the same year. This baseline is specifically tailored to each source country and year, representing the actual distribution of citations accumulated over a 5-year window from the source country to all global publications within that year. The resulting measure of citation preference between a source and target country can be interpreted as the probability that a randomly selected publication from the target country has more citations from the source country than a randomly selected publication from anywhere else in the world, and assumes a value between 0 (strong preference against) and 1 (strong preference for), where a value of 0.5 captures no preference. Since our method aggregates over a 5-year citation window, the most recent year for our analysis is 2017.

There are many possible mechanisms that may contribute to strong citation preferences; our data lets us further control for two potential contributions. First, scientists are known to self-cite (Aksnes, 2003) at rates which vary based on culture,

discipline, and demographics (King et al., 2017; Azoulay and Lynn, 2020). Second, sharing an affiliation can increase the propensity to engage with a scientist's work (Wuestman et al., 2019). To control for the possible influence of these two factors, we removed all citations between publications that share at least one author or at least one affiliation (SI, Section S2). This framework further accommodates controlling for specific factors which may influence national citation preferences, including scientific disciplines and journals, by modifying the citation baseline (see Methods).

Data and Methods

Bibliometric Data

The dataset was drawn from the OpenAlex (Priem et al., 2022) bibliometric database in July 2022. OpenAlex is built upon the Microsoft Academic Graph (MAG), which Microsoft shuttered in December 2021, CrossRef, and ORCID. We used all indexed "journal-article" and "proceedings-article" records listed as published after 1990 and excluded any publication that did not list an institutional address.

Publications are associated with countries using the institutional addresses listed by the authors. We assign a full unit credit of a publication to every country of affiliation on the paper's author byline ("full counting"). For example, a paper listing ten authors– three with affiliations in Hungary, five with affiliations in the United States, and two in Canada— would count one paper to all three countries. See Supplementary Information for more details.

National Co-variate Data

We use data on national GDP, GDP per capita, and Population from the World Bank (Fantom and Serajuddin, 2016) to approximate the economic wealth and size of each country. The dataset covers 264 countries from 1960 to 2023. The official spoken language is provided for 195 countries and is encoded as a binary variable denoting common language for country pairs (Melitz and Toubal, 2014). We also source the bilateral distances (in kilometres) for most country pairs across the world from the *GeoDist* dataset provided by the Centre for Prospective Studies and International Information (CEPII) (Mayer and Zignago, 2011). This dataset also provides the continent each country belongs to, which we convert into a binary indicator denoting whether two countries belong to the same continent. In addition, Science and Technology Agreements (STA) are regarded as an important tool to achieve strategic Science Diplomacy (SD) objectives (Langenhove, 2017). We select records of STAs between countries (Nicolas Ruffin and¨ Schreiterer, 2017) to obtain the cumulative number of STAs between two countries over time.

National citation preference

We fix a year y and a source country (citing country) s and identify all publications with at least one affiliation in the source country over the next 5 years (y to y+5). We then find all publications worldwide published in year y that also received citations

from the source country's 5-year publications. This process generates country-specific citation frequencies ($c_{s,5}$) over the fiveyear observation window, enabling us to establish a hierarchical ranking of $n_{s,y}$ publications that have garnered at least one citation from the source country ($c_{s,5} >= 1$). This forms the baseline sample, comprising a citation distribution $p(c_5|s,y)$ specific to the source country s and year y, with a sample size of $n_{s,y}$. Next, we narrow our analytical focus to a designated target country t, identifying a subset of $n_{s,t,y}$ publications within our sample. These publications, represented by the distribution $p(c_5|s,t,y)$, must satisfy two criteria: they have received citations from the source country and maintain at least one institutional affiliation within the target country.

The national citation preference, $P_{s,t,y}$, from the source country s to the target country t in year y is found using the Area Under the receiver-operator Curve (AUC) as a measure of the extent to which the target country's publications are randomly distributed throughout the source country's ranking. Specifically, the national preference is found as:

$$\mathrm{P_{s,t,y}} = \frac{1}{\mathrm{n_{s,t,y}\, n_{s,y}}} \sum_{\mathrm{i=1}}^{\mathrm{n_{s,t,y}}} \sum_{\mathrm{j=1}}^{\mathrm{n_{s,y}}} \mathbb{I}\left(\mathrm{c_5^{(i)}} > \mathrm{c_5^{(j)}}\right) \quad (1)$$

where $c_5^{(i)}$ is the i-th sample from $p(c_5|s,t,y)$, $c_5^{(j)}$ is the the j-th sample from $p(c_5|s,y)$, and $\mathbb{I}$ is the indicator function, which is 1 if $c_5^{(i)} > c_5^{(j)}$ and 0 otherwise. The AUC is a measure of the probability (between 0 and 1) that a randomly chosen publication from the cited country is ranked higher than a randomly chosen publication from any other country; a value of 1 reflects the cited country's publications are over-expressed towards the top of the ranking, 0 occurs when the cited country's publications are under-expressed towards the bottom of the ranking, and 0.5 denotes a random distribution throughout the ranking.

We can further quantify the statistical significance of the over/under-representation of a specific country in the citation counts due to the equivalence of the AUC and Mann-Whitney U statistic (a.k.a. the Wilcox rank sum statistic). Specifically, we follow DeLong et al. to compare the observed AUC to 0.5 (DeLong et al., 1988) using the algorithm's fast implementation (Sun and Xu, 2014).

International citation preference network

The international citation preference network is a temporal network, independently constructed for each year. To avoid multiple hypothesis testing, we used the Holm step-down method (Holm, 1979) using Bonferroni adjustments as implemented in Statsmodels with $\alpha = 0.01$. The cumulative network aggregates over of all time slices and adopts the sign of the most recent slice in which the edge appeared.

The community structure within the positive international citation preference network is found using the Degree Corrected Stochastic Block Model (DCSBM) as implemented in graphtool (Peixoto, 2017). Network centrality for the positive international citation preference network is found using the PageRank algorithm with a return probability of $\alpha = 0.85$.

Stratified bootstrap baseline

To account for potential explanatory factors such as disciplinary focus and journal quality, we refine the assumptions underlying the random baseline in our national citation preference measure. We achieve this by implementing a stratified bootstrap approach, where we sample from the conditional citation distribution while ensuring that the sampled set exactly matches the observed publication counts for each journal in the observed citation distribution. Specifically, given the sample of $n_{s,t,y}$ publications affiliated with the target country t in year y and cited by the source country s, we track the frequency with which each journal appears, denoted $j_{s,t,y}$. We then sample with replacement from the source country's baseline distribution $p(c_5|s,y)$ such that the journal counts remain consistent with the observed values. This adjustment controls for the influence of journal-specific factors and disciplinary differences. We then perform 100 samples of this bootstrap procedure and use the mean and standard deviation of the AUCs to identify statistically significant links.

Scientific ideas

To identify scientific ideas, we follow the methodology introduced in Cheng et al. 2023 (Cheng et al., 2023). Specifically, we analyze the titles and abstracts for all of the publications in our OpenAlex corpus to identify the publications that mention at least one of 46,535 scientific ideas derived by Cheng et al. using the data-driven phrase segmentation algorithm, AutoPhrase (Shang et al., 2018). We then post-process these ideas, removing cases that were first mentioned before 2000 and focusing only on those ideas that were mentioned by only one country in their first year of usage, resulting in 7,327 unique ideas mentioned in 202,932 publications. Finally, we derive a dyadic variable, for all pairs of countries in our network that mentioned at least one idea, denoting the fraction of ideas whose first usage was in the Origin country and then were later used in the Destination country.

Logistic regression analysis

We use a logistic regression model to investigate the potential relationship between the propensity for scientific ideas to spread between countries and their connectivity in the international citation preference network. The model is written as follows:

$$\log \quad \frac{y_c}{1-y_c} = \beta_0 + \beta_1 X_{1,c} + \beta_2 X_{2,c} + \cdots + \beta_k X_{k,c} \qquad (2)$$

Where c denotes countries and y_c is the dependent variable. For the first group of models, we use the fraction of ideas that originate in the origin country and are later mentioned by the destination country (see Methods and SI, section S2). The included control variables are the GDP and Population for both the Origin and Destination countries. The investigated independent variables are the total number of ideas mentioned by the Origin and Destination countries, the Topical Distance between the countries' publications, the Physical Distance between the countries, a binary indicator of common official language, the one-hot encoding of a directed positive edge from the Destination to the Origin in the international citation preference

network, the onehot encoding of a directed negative edge from the Destination to the Origin in the international citation network, and the PageRank centrality of the Origin and Destination countries in the positive international citation preference network. We apply log-transformation with base 10 to GDP, Population, and Physical Distance. All features besides the binary features (Same Official Lang, Positive Edge, Negative Edge) are standarized by subtracting the mean and dividing by the standard deviation.

Fixed-effect multinomial logistic regression

We use the multinomial logit model to predict the trinary citation preference between countries (e.g. positive, negative, or no preference). The multinomial logit model assumes that the log odds of each category $s \in \{-1,1\}$ relative to the reference category of no citation preference ($s = 0$) is a linear combination of the independent variables. Specifically, the model is defined as follows:

$$\log\left(\frac{P(Y_{ijt}=s)}{P(Y_{ijt}=0)}\right) = \beta_{s0} + \beta_{s1}X_{it} + \beta_{s2}X_{jt} + \beta_{s3}X_{ijt} + \alpha_t \quad (3)$$

where $P(Y_{ijt} = s)$ is the probability of the edge sign between source country i and target country j at time t taking value $s \in \{-1,1\}$; X_{it} and X_{it} capture potential country-specific characteristics in the country i and j at time t, respectively, while X_{ijt} represents potential pair-specific barriers or catalysts between country i and j at time t; α_t are the time-specific effects (intercepts) that capture the heterogeneity across time periods. β_{s0} is the intercept for category s; β_{s1}, β_{s2} and β_{s3} are the coefficients associated with the independent variables $X_i, X_j$and X_{ijt} for category s. We investigate different variants of the above model to study different combinations of countryspecific and country-pair-specific variables. The included control variables are the GDP per capita, population, and the fraction of top journal publications for both the Source and Target countries. The investigated pair-specific independent variables are physical distance, field distance, the same continent, the same official language, the cumulative number of bilateral science and technology agreements and scientific collaboration strength. We apply log-transformation with a base 10 to GDP per capita, population, physical distance, the cumulative number of bilateral science and technology agreements and scientific collaboration strength. All non-binary features are standardized by subtracting the mean and dividing by the standard deviation.

Results

International network of scientific recognition

We first build the network of international scientific recognition (Fig. 1A). The international scientific recognition network is a temporal signed and directed network in which each country is a node, and a source country is linked to a target country by a positive (negative) edge if the source country over-cites (under-cites)

the target country's publications. To begin, we consider the cumulative network in which we aggregate over time, taking any edge that appears at least one throughout the 27 years. We find that 147 countries had at least one statistically significant relationship to be included in the network. Of the 17,030 possible international relationships, only 541 are positive interactions and 1538 are negative interactions. The positive recognition network is shown in Fig. 1B. Scientific publications from Switzerland are over-cited the most, with 36 incoming edges, followed by Great Britain, Germany, and the Netherlands (Fig. 2B). On the other hand, publications from China are the most under-cited, with 86 incoming undercitation edges, followed by Japan, Iran and India (Fig. 2C).

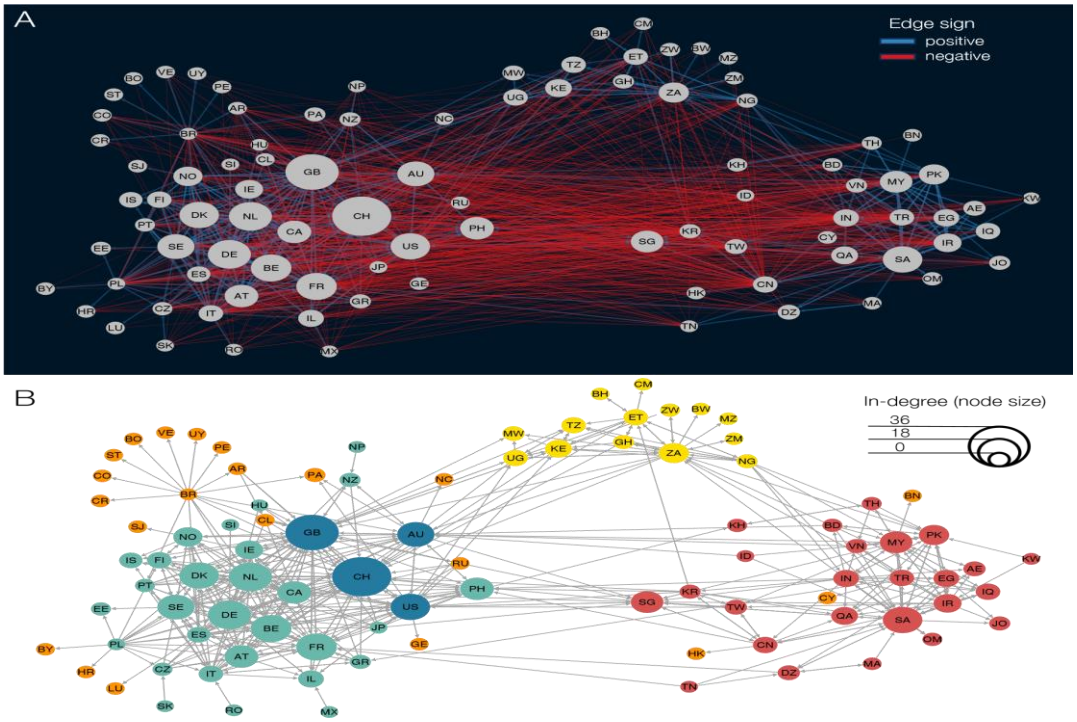


Figure 1. International network of scientific citation preferences. A) The full international network of scientific citation preferences. Edge color reflects positive (blue) or negative (red) citation preferences. B) The network filtered to positive relationships. The node size captures the country in-degree, while node colour reflects membership in one of five communities inferred using the degree-corrected stochastic block model. Node position is the same in both panels and was derived using only the positive relationships.

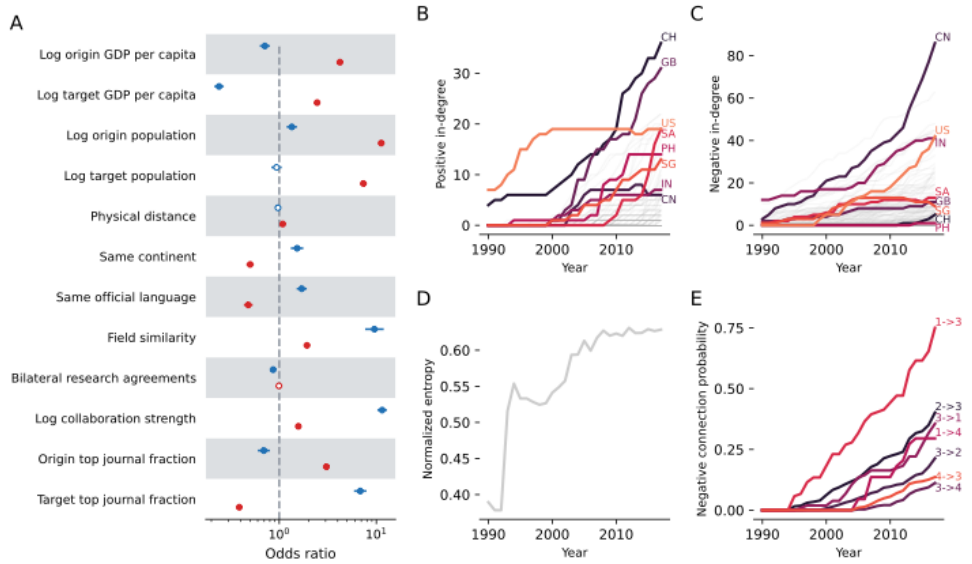


Figure 2. Properties of the international network of scientific citation preferences. (A) The odds ratios for a multinomial logit model with temporal fixed-effects to predict the positive (blue) or negative (red) citation preference compared to the baseline of no preference. Solid points are statistically significant at $p < 0.05$ with the 95% confidence intervals shown. The full regression table can be found in the SI, Table S3. The (B) positive and (C) negative in-degrees highlight 6 prominent countries, including the most positively viewed country in 2017, Switzerland (CH), and the most negatively viewed country, China (CN). (D) The normalized entropy for the distribution of PageRank centrality over the nodes has been increasing over the last 30 years. (E) The probability for a negative citation preference between a country in a source community and a country in a target community.

To identify key country-specific and country-pair-specific factors related to national citation preferences, we build a multinomial logit model with temporal fixed-effects to predict the citation preference between all pairs of countries from 1990 through 2017. We find that, while most independent variables play a statistically significant role in the prediction, many of them do not differentiate in terms of the contribution direction between positive and negative citation preferences (Fig. 2A and SI, Table S1). In particular, collaboration strength, while indicative of a link between countries, does not help differentiate the sign of that preference, and topical similarity only contributes to the prediction of positive preferences. However, three cultural indicators: common language ($\beta_{positive} = 0.53$, 95% $CI = [0.41, 0.65]$; $\beta_{negative} = -0.74$, 95% $CI = [-0.84, -0.63]$), same continent ($\beta_{positive} = 0.42$, 95% $CI = [0.27, 0.57]$; $\beta_{negative} = -0.69$, 95% $CI = [-0.78, -0.6]$), and participation in Science and Technology Agreements

(bilateral research agreement, $\beta_{positive} = -0.17$, 95% $CI = [-0.19, -0.15]$; $\beta_{negative} = -0.01$, 95% $CI = [-0.02, 0.0]$), relate to both the presence and sign of the national citation preference (Fig. 2A). Finally, we use the fraction of publications in top journals to capture one aspect of research quality (see Methods and SI, Section S1)

and find that higher-quality publications in the target country are associated with a higher probability of a positive citation preference from the origin country, while lower-quality publications in the target country are associated with a negative citation preference ($\beta_{positive}$ = 1.51, 95% CI = [1.39,1.63]; $\beta_{negative}$ = −0.75, 95%CI =[−0.8,−0.69]). Overall, this model suggests a mutual-influence relationship between scientific quality, national culture, science diplomacy and international scientific recognition.

Mapping the network of international preferences over time reveals the changing landscape of scientific diplomacy. Specifically, the network of international citation preferences has evolved away from a core-periphery structure dominated by a few hubs to a more distributed structure, a change which we measure by the increasing normalized entropy for the distribution of PageRank centrality scores (Fig. 2D). For example, before 2000, the network was dominated by the United States, with relatively little positive scientific recognition of countries in Asia or Africa (Fig. 2B). However, by 2010, Switzerland and Great Britain surpassed the United States in global recognition, and there were notable rises in recognition to Saudi Arabia, the Philippines, and Singapore (Fig. 2B). Throughout this period, China and Japan remained the most under-cited, dominating the negative citation preference network (Fig. 2C).

Growing international scientific fragmentation

The preference of some nations for the scientific work of others, combined with the proliferation of negative biases against groups of countries, is a characteristic hallmark of international scientific fragmentation (Aref et al., 2020). This pattern in citation patterns can stem from various factors, such as disciplinary biases, prevailing research trends, language barriers, geographical disparities, or ideological preferences. As a result, scientific fragmentation can distort the perception of research's importance and impact, reinforce existing knowledge gaps, and impede the equitable dissemination and recognition of diverse scientific contributions.

To measure the dynamics of international scientific fragmentation, we first detect the presence of international communities using the degree-corrected stochastic block model, finding strong evidence for a partition of the positive network in 5 distinct communities. Three blocks strongly resemble a three-layer core-periphery structure (Gallagher et al., 2021). Specifically, we find a dense core consisting of Western countries that tend to positively prefer each other's work (1, dark blue) and a weaker core of many European countries (2, light blue), while countries in the periphery (5, orange) are agnostic to each other, but prefer countries from both the weak and strong cores (Fig. 1).

At the same time, this analysis confirms that the core-periphery structure is an oversimplification of the diverse communities in global science. The international scientific recognition network reveals two additional communities outside of the Western scientific world: one community captures an international community predominately composed of Asian countries (3, red), including both East Asia and the Middle East, while another reflects the African nations (4, yellow).

The fragmentation of global science is evidenced by the distribution pattern of positive and negative citation preferences across scientific communities. Overall, only 34% of positive citation preferences occur between nations from different communities, whereas negative citation preferences predominantly cross community boundaries, with over 86% occurring between nations from different communities.

The structure of the international citation preference network and its communities provides a more nuanced view of the differing roles nations play in shaping global scientific recognition and knowledge dissemination. For example, while both Singapore and China have gained recognition for their scientific contributions (Zhou and Leydesdorff, 2006), our analysis shows that Singapore occupies a unique bridging role between different communities, whereas China, despite its prominence, remains within the Asian community without holding a central core position. Notably, our work highlights Saudi Arabia, Turkey, and Iran as occupying more central roles within the Asian scientific community. Similarly, South Africa (ZA) stands out as a central node within the African scientific community, while the network reveals the distinct roles of Uganda and Nigeria as key bridges—Uganda connecting to Western communities and Nigeria to the Asian community.

To assess the dynamics of international scientific fragmentation, we look at the probability of forming negative or positive links. Overall, we observe a growing tendency for nations to negatively judge the work of other nations as evidenced by the increase in negative connection probabilities (SI, Figure S1). However, the community structure of the international scientific recognition network reveals that these preferences are not evenly distributed and are not primarily directed at specific nations. Instead, the fragmentation of global science appears to be influenced by the detected geopolitical communities. As shown in Fig. 2E, the probability of inter-community negative preference links has grown significantly since 1990. The probability of negative inter-community links is largest between the Western and Asian communities, specifically communities $1 -> 3$ and $3 -> 1$ as well as $2 -> 3$ and $3 -> 2$, but has also significantly grown between the African and Western communities $1 -> 4$, $4 -> 1$ and the African and Asian communities $3 -> 4$, $4 -> 3$. Significantly, there are nearly symmetric negative inter-community link probabilities, indicating the true fragmentation of the global scientific landscape into distinct communities cannot be explained by a core-periphery model.

International recognition network and the diffusion of ideas

To explore the potential connection between the position of nations in the international scientific recognition network and the propensity for them to spread ideas, we investigate the flow of knowledge between countries. We operationalize scientific knowledge through the appearance of keywords in the title and abstract of scientific publications (Milojevic et al.´ , 2011; Milojevic´, 2015; Cheng et al., 2023). Specifically, we identify the mention of over 40,000 n-grams defined as scientific ideas by a previous study (Cheng et al., 2023) and limit to 7,327 unique ideas originating in only one country after 2000 (see Methods and SI, Section S2). We then model the fraction of ideas originating in one country that are eventually mentioned

in another target country at least once during the subsequent 22 years (2000-2022) using logistic regression. This approach allows us to gauge the spread of information through the global scientific ecosystem, reflecting the broader exchange of ideas without needing to follow each idea's continuous trajectory over time. Consequently, we use the cumulative international recognition network where we aggregate into a static snapshot using all links that appear in at least one time slice.

Since there are many factors which may influence the flow of knowledge between countries, in Model 1, we predict the fraction of ideas which spread between 9,635 country pairs based on the number of ideas which originate and terminate in the origin and target countries respectively, the countries' populations and GDP per capita. Unsurprisingly, the model coefficients in Table 1 show that the number of ideas originating in a country, the ability of a target country to take up ideas, and the country's population are all statistically significant. We also find that the topical distance between the countries' scientific publications and whether the origin and destination share a common language are also statistically significant in their relation to the spread of ideas.

Table 1. International diffusion of scientific ideas. Model coefficients for a series of logistic regression models to predict the fraction of scientific ideas that originate in one country that are eventually mentioned in the destination country. Confidence intervals in parentheses. Standard errors and p-values are reported.

	Dependent variable: Fraction of scientific ideas.			
	Model			
	(1)	(2)	(3)	(4)
Intercept	−1.08***	−1.1***	−1.08***	−1.08***
	(−1.13,−1.02)	(−1.16,−1.04)	(−1.15, −1.02)	(−1.15, −1.01)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0
Log Population origin	−0.38***	−0.45***	−0.44***	−0.46***
	(−0.44,−0.33)	(−0.51,−0.38)	(−0.51,−0.37)	(−0.53,−0.39)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.04; p-v 0.0
Log Population destination	0.17**	0.18**	0.17**	0.18**
	(0.06,0.28)	(0.06,0.3)	(0.06,0.29)	(0.06,0.3)
	S.E. 0.06; p-v 0.0019	S.E. 0.06; p-v 0.0036	S.E. 0.06; p-v 0.0042	S.E. 0.06; p-v 0.0034
Log GDP per capita origin	−0.14***	−0.21***	−0.21***	−0.25***
	(−0.2,−0.09)	(−0.27,−0.15)	(−0.28,−0.15)	(−0.32,−0.18)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.04; p-v 0.0
Log GDP per capita destination	0.07	0.07	0.07	0.07
	(−0.05,0.19)	(−0.06,0.2)	(−0.06,0.2)	(−0.06,0.2)
	S.E. 0.06; p-v 0.2372	S.E. 0.06; p-v 0.2687	S.E. 0.07; p-v 0.2719	S.E. 0.07; p-v 0.2703
Number of ideas origin	0.16***	0.19***	0.19***	0.18***
	(0.1,0.21)	(0.13,0.25)	(0.12,0.25)	(0.11,0.24)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0
Number of ideas destination	0.81***	0.82***	0.83***	0.83***
	(0.69,0.94)	(0.68,0.96)	(0.69,0.97)	(0.68,0.97)
	S.E. 0.06; p-v 0.0	S.E. 0.07; p-v 0.0	S.E. 0.07; p-v 0.0	S.E. 0.08; p-v 0.0

Topic distance		−0.1** (−0.17,−0.03) S.E. 0.03; p-v 0.0046	−0.09** (−0.16,−0.03) S.E. 0.03; p-v 0.007	−0.09** (−0.16,−0.02) S.E. 0.03; p-v 0.0084
Log Physical distance		−0.01 (−0.07,0.05) S.E. 0.03; p-v 0.6742	0.01 (−0.05,0.07) S.E. 0.03; p-v 0.6855	0.01 (−0.05,0.07) S.E. 0.03; p-v 0.811
Common language		0.39*** (0.21,0.57) S.E. 0.09; p-v 0.0	0.34*** (0.15,0.53) S.E. 0.1; p-v 0.0004	0.32*** (0.13,0.51) S.E. 0.1; p-v 0.0008
Positive citation preference			0.4** (0.16,0.64) S.E. 0.12; p-v 0.001	0.32* (0.07,0.57) S.E. 0.13; p-v 0.0111
Negative citation preference			−0.23* (−0.41,−0.05) S.E. 0.09; p-v 0.0114	−0.22* (−0.4,−0.04) S.E. 0.09; p-v 0.0153
Network centrality origin				0.08* (0.01,0.16) S.E. 0.04; p-v 0.0218
Network centrality destination				0.01 (−0.05,0.08) S.E. 0.03; p-v 0.6583
Note:		*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$		
Observations	9635	8292	8292	8292
Pseudo-R^2	0.1652	0.1876	0.1924	0.1932
Log Likelihood	-4242.81	-3606.07	-3584.55	-3581.06
F statistic	243.97*** (d.f.=6.0)	148.93*** (d.f.=9.0)	123.34*** (d.f.=11.0)	104.5*** (d.f.=13.0)

The network of scientific recognition enhances our ability to predict the flow of ideas between countries, as shown in Models 3 and 4 (Table 1). The odds ratios suggest that a positive recognition edge between the target and originating countries leads to a 1.5 times increase in the fraction of ideas which spread between those countries compared to the baseline of no edge, while a negative recognition edge between the target and originating countries leads to 0.8x decrease in the fraction of ideas which spread between those countries. Beyond the immediate neighborhood, the global network topology is hypothesized to play a significant role in the spread of information over social networks (Kempe et al., 2005; Pei et al., 2018). We also find that the network centrality of the originating country is related to the diffusion of ideas (*p-value* < 0.0218; 95% IC = [0.01,0.16]) (see Table 1 for details).

Exploring the impact of journals on citation preferences

We now extend our analysis by introducing additional controls to further explore factors influencing citation preferences. Our framework seamlessly integrates a non-parametric approach that accounts for the field or journal in which each article is published, allowing us to control for variability in citation practices across disciplines and venues. By incorporating these controls and juxtaposing the new network against our original, this enhanced model provides a more refined

understanding of how disciplinary and journal-specific effects interact with national-level citation behaviors, offering deeper insights into the structure of global scientific recognition.

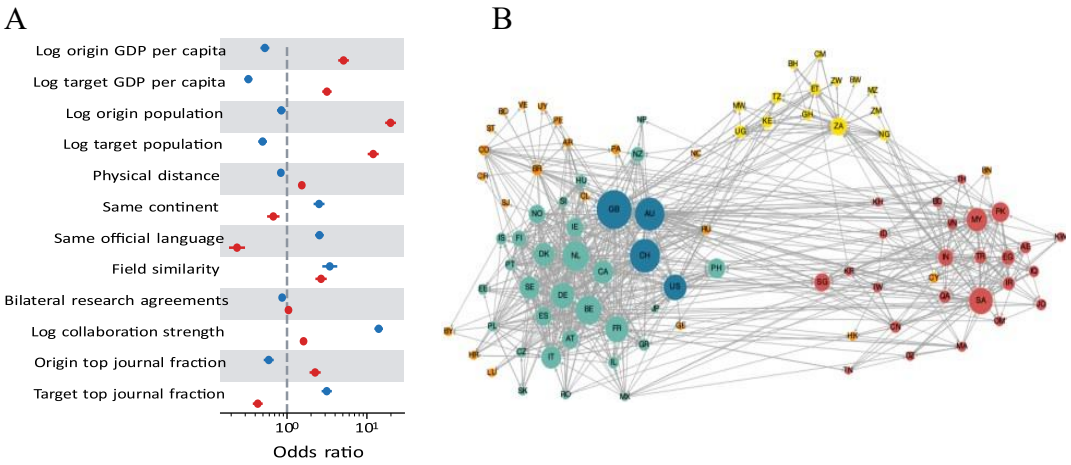


Figure 3. The international network of scientific citation preferences controlling for publication journal. (A) The odds ratios for a multinomial logit model to predict the positive (blue) or negative (red) citation preference compared to the baseline of no preference. Solid points are statistically significant at $p < 0.05$ with the 95% confidence intervals shown. The full regression table can be found in the SI, Table S4. (B) The journal bootstrap network filtered to positive relationships using the same layout as in Fig 1.

Instead of relying on the full citation distribution for all publications cited by the source country, we construct a new baseline citation distribution using a stratified bootstrap approach that accounts for journal frequency (see Methods for details). This technique samples from the source country's conditional citation distribution while ensuring the sampled set reflects the observed publication counts for each journal. By controlling for journal-level citation patterns– commonly used as proxies for scientific discipline and "quality"–this method provides a more detailed benchmark, isolating national citation preferences from journal-related con-founders. Shown in Fig. 3B, the resulting cumulative international network of citation preferences based on the journal bootstrap (N2) exhibits both notable similarities and differences when compared to the original network (N1). Specifically, N2 reveals more positive national preferences, with a total of 645 compared to 541 in N1, while it shows significantly fewer negative preferences, dropping from 1,538 in N1 to just 334 in N2. At the same time, there is considerable overlap between the networks: 448 positive preferences are present in both networks, accounting for 84% of the smaller N2, and 326 negative preferences are shared, representing 98% of the smaller N1. The variation in positive edges is largely concentrated in a small number of countries: 47% of the new edges are directed toward just 11 countries, while 30% originate from only 7 countries. Moreover, the edge distribution in N2 largely mirrors

the community structure observed in N1 such that 60% of positive edges connect nations within the same community in N2, slightly down from 66% in N1, and 85% of negative edges link nations from different communities in N2, compared to 86% in N1. Using a similar multinomial logistic regression model with temperal fixed-effects to predict the presence and sign of national preferences, we find the same independent variables play remarkably similar patterns of importance for predicting the odds of a positive or negative edge, and differentiating between those signs (Fig. 3A).

Taken together, these observations suggest that about 80% of the negative citation preferences we initially identified can be attributed to disciplinary differences in scientific focus and journal "quality". At the same time, the increase in positive preferences primarily within the original communities, indicates the importance of those communities, suggesting they are highly influential in shaping collaborative networks and recognition. Ultimately, these findings emphasize the value of applying robust methodological frameworks to uncover the complexities of international citation preferences, providing deeper insights into the factors that influence scientific recognition on a global scale.

Discussion

The international scientific landscape, a complex and dynamic web of knowledge, people and practices, is molded by national interests grounded in historical events, cultural values, political agendas, economics, and technological innovations. These same forces shape interactions between nations through incentives for international collaboration, researcher mobility, and knowledge flows. By analyzing more than fifty-seven million scientific publications across 223 countries spanning the period 1990-2022, we provide a large-scale temporal and structural analysis of the collective structure of global scientific recognition. We find that the international citation preference network constructed from these publications is shaped by cultural elements, including language and political agreements, and augments insights from the study of scientific collaboration and scientific topics. Additionally, we quantify the network's departure from a core-periphery structure and identify five communities corresponding to major global regions, revealing a growing trend towards increased fragmentation. We then demonstrate that the international citation preference network imposes limitations on the dissemination of scientific ideas, reflecting a more efficient spread of concepts within a community compared to their transmission between distinct communities. Finally, we find that around 80% of negative citation preferences can be explained by disciplinary differences and journal "quality", while those same factors increase the prevalence of positive preferences within the original communities.

Our results reveal the collective structure of international citation preferences, complementing the viewpoints offered by collaboration, migration, and citation volume (Glanzel¨, 2001; Leydesdorff and Wagner, 2008; Wagner and Jonkers, 2017). While we were able to quantify the magnitude and significance of these preferences, and mapped how these preferences changed when controlling for scientific journals,

but our data is unable to suggest all of causal mechanisms driving them. Additional work is needed to differentiate whether the observed patterns are rooted in social factors like cultural differences or accessibility. Nevertheless, the resulting model of global recognition reveals interesting features of the international state of scientific discourse.

Our quantitative results reveal the emergence of multiple distinct international communities, challenging the traditional core-periphery model of global science. These findings show that rather than a simple transition of countries from the periphery to the core, regional scientific communities are increasingly disconnected from each other. Notably, we identified negative citation links between these communities–evidence of declining mutual recognition–which would not be captured by a standard citation or collaboration network model. This suggests that these communities are growing apart, reinforcing their preference for internal knowledge sharing over external engagement. The implications of this are profound for the sociology of science and global science inequalities: instead of a unified global hierarchy, we may be witnessing the fragmentation of scientific influence, where certain regions strengthen internal ties at the cost of broader visibility and integration into the global scientific landscape. This deepens existing disparities, as regions that were once peripheral may develop more insular networks, further complicating efforts to address global inequalities in scientific recognition and collaboration.

The results of our study on the international scientific landscape carry several policy implications. Firstly, acknowledging the influence of national interests, historical events, cultural values, political agendas, economics, and technological innovations on global scientific recognition through citation suggests that the assessment of scientific impact to publications, authors, and organizations should also be sensitive to these multifaceted factors. It further suggests research into the implications of national vs international citation recognition on individual careers and potential inequalities in recognition that may arise (Huang et al., 2020). Secondly, the departure from a traditional core-periphery structure via the identification of five major global communities, underscores a growing trend towards increased fragmentation in scientific influence. Policymakers will need to adapt strategies to address this shift, ensuring inclusivity and collaboration across diverse scientific communities. Finally, the negative influences on national preferences of bilateral agreements for science and technology mirror results found for other types of treaties (Hoffman et al., 2022). This finding underscores that such agreements, which are intended to foster collaboration and knowledge exchange between nations, may encounter challenges that impede their effectiveness. Thus, they hamper an important tool that policymakers have to establish and nurture international scientific relationships, potentially hindering the full realization of the intended benefits of bilateral agreements in advancing global scientific cooperation.

Code and Data Availability

The primary dataset, OpenAlex, is freely available online at https://openalex.org/. All code used to conduct the analysis and generate the figures, as well as the processed data and network structure, is included as part of the pySciSci Python package (Gates and Barabasi´, 2023):
https://github.com/SciSciCollective/pyscisci/globalscience.

References

Abramo, G., C. A. D'Angelo, and F. Di Costa (2020, February). The role of geographical proximity in knowledge diffusion, measured by citations to scientific literature. *Journal of Informetrics 14*(1), 101010.

Adams, J. (1998). Benchmarking international research. *Nature 396*(6712), 615–618.

Adams, J. (2012). The rise of research networks. *Nature 490*(7420), 335–336.

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics 56*(2), 235–246.

Althouse, B. M., J. D. West, C. T. Bergstrom, and T. Bergstrom (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology 60*(1), 27–34.

Anderson, W. (2018). Remembering the spread of western science. *Historical Records of Australian Science 29*(2), 73–81.

Aref, S., L. Dinh, R. Rezapour, and J. Diesner (2020). Multilevel structural evaluation of signed directed social networks based on balance theory. *Scientific Reports 10*(1), 15228.

Azoulay, P. and F. B. Lynn (2020). Self-citation, cumulative advantage, and gender inequality in science. *Sociological Science 7*, 152–186.

Basu, A., P. Foland, G. Holdridge, and R. D. Shelton (2018, October). China's rising leadership in science and technology: quantitative and qualitative indicators. *Scientometrics 117*(1), 249–269.

Bauder, H., O. Lujan, and C.-A. Hannan (2018). Internationally mobile academics: Hierarchies, hegemony, and the geo-scientific imagination. *Geoforum 89*, 52–59.

Cheng, M., D. S. Smith, X. Ren, H. Cao, S. Smith, and D. A. McFarland (2023, April). How new ideas diffuse in science. *American Sociological Review*, 000312242311669.

Chinchilla-Rodr´ıguez, Z., C. R. Sugimoto, and V. Lariviere (2019, June). Follow the leader:` On the relationship between leadership and scholarly impact in international collaborations. *PLoS One 14*(6), e0218309.

Choi, S. (2012, January). Core-periphery, new clusters, or rising stars?: international scientific collaboration among 'advanced' countries in the era of globalization. *Scientometrics 90*(1), 25–41.

Collyer, F. (2014, September). Sociology, sociologists and core-periphery reflections. *Journal of Sociology 50*(3), 252–268.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988, Sep). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics 44*(3), 837.

Fantom, N. and U. Serajuddin (2016). *The World Bank's classification of countries by income*. The World Bank.

Freeman, R. B. (2010, July). Globalization of scientific and engineering talent: international mobility of students, workers, and ideas and the world economy. *Economics of Innovation and New Technology 19*(5), 393–406.

Gallagher, R. J., J.-G. Young, and B. F. Welles (2021, March). A clarified typology of coreperiphery structure in networks. *Science Advances 7*(12), eabc9800.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Jama 295*(1), 90–93.

Gates, A. J. and A.-L. Barabasi (2023). Reproducible science of science at scale:´ *pySciSci*. *Quantitative Science Studies*, 1–17.

Glanzel, W. (2001). National characteristics in international scientific co-authorship relations.¨ *Scientometrics 51*(1), 69–115.

Gomez, C. J., A. C. Herman, and P. Parigi (2022). Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour 6*(7), 919–929.

Gong, K., J. Xie, Y. Cheng, V. Lariviere, and C. R. Sugimoto (2019). The citation advantage` of foreign language references for Chinese social science papers. *Scientometrics 120*(3), 1439–1460.

Gui, Q., C. Liu, and D. Du (2019). Globalization of science and international scientific collaboration: A network perspective. *Geoforum 105*, 1–12.

Galvez, A., M. Maqueda, M. Martinez-Bueno, and E. Valdivia (2000). Scientific publication´ trends and the developing world: What can the volume and authorship of scientific articles tell us about scientific progress in various regions? *American Scientist 88*(4), 526–533.

Hagstrom, W. O. (1974). Competition in science. *American Sociological Review 39*(1), 1–18.

Heimeriks, G. and R. Boschma (2014, March). The path- and place-dependent nature of scientific knowledge production in biotech 1986-2008. *Journal of Economic Geography 14*(2), 339–364.

Hoffman, S. J., P. Baral, S. Rogers Van Katwyk, L. Sritharan, M. Hughsam, H. Randhawa,
G. Lin, S. Campbell, B. Campus, M. Dantas, N. Foroughian, G. Groux, E. Gunn, G. Guyatt, R. Habibi, M. Karabit, A. Karir, K. Kruja, J. N. Lavis, O. Lee, B. Li, R. Nagi, K. Naicker, J.A. Røttingen, N. Sahar, A. Srivastava, A. Tejpar, M. Tran, Y.-q. Zhang, Q. Zhou, and M. J. P. Poirier (2022, August). International treaties have mostly failed to produce their intended effects. *Proceedings of the National Academy of Sciences 119*(32), e2122854119.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Huang, J., A. J. Gates, R. Sinatra, and A.-L. Barabasi (2020). Historical comparison of gender´ inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences 117*(9), 4609–4616.

Kempe, D., J. Kleinberg, and E. Tardos (2005). Influential nodes in a diffusion model for so-´ cial networks. In *Automata, Languages and Programming: 32nd International Colloquium, ICALP 2005, Lisbon, Portugal, July 11-15, 2005. Proceedings 32*, pp. 1127–1138. Springer.

King, D. A. (2004). The scientific impact of nations. *Nature 430*(6997), 311–316.

King, M. M., C. T. Bergstrom, S. J. Correll, J. Jacquet, and J. D. West (2017, Dec). Men set their own cites high: Gender and self-citation across fields and over time. *Socius 3*, 2378023117738903.

Kozlowski, D., V. Lariviere, C. R. Sugimoto, and T. Monroe-White (2022, January). Inter-` sectional inequalities in science. *Proceedings of the National Academy of Sciences 119*(2), e2113067119.

Langenhove, L. V. (2017). Tools for an eu science diplomacy. *Luxembourg: European Commission*.

Lariviere, V., K. Gong, and C. R. Sugimoto (2018). Citations strength begins at home. *Nature 564*(7735), S70–S70.

Leydesdorff, L., C. Wagner, H. W. Park, and J. Adams (2013). International collaboration in science: The global map and the network. *arXiv preprint arXiv:1301.0801*.

Leydesdorff, L. and C. S. Wagner (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics 2*(4), 317–325.

Marginson, S. (2022a, August). What drives global science? The four competing narratives. *Studies in Higher Education 47*(8), 1566–1584.

Marginson, S. (2022b). What drives global science? The four competing narratives. *Studies in Higher Education 47*(8), 1566–1584.

Marginson, S. and X. Xu (2023). Hegemony and inequality in global science: Problems of the center-periphery model. *Comparative Education Review 67*(1), 31–52.

May, R. M. (1997). The scientific wealth of nations. *Science 275*(5301), 793–796.

Mayer, T. and S. Zignago (2011). Notes on CEPII's distances measures: The geodist database. *CEPII Documentation*.

McKiernan, E. C., L. A. Schimanski, C. Munoz Nieves, L. Matthias, M. T. Niles, and J. P.˜ Alperin (2019). Use of the journal impact factor in academic review, promotion, and tenure evaluations. *Elife 8*, e47338.

Melitz, J. and F. Toubal (2014). Native language, spoken language, translation and trade. *Journal of International Economics 93*(2), 351–363.

Milojevic, S. (2015, October). Quantifying the cognitive extent of science.´ *Journal of Informetrics 9*(4), 962–973.

Milojevic, S., C. R. Sugimoto, E. Yan, and Y. Ding (2011, October). The cognitive structure´ of Library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology 62*(10), 1933–1953.

Moravcsik, M. J. (1985, March). Applied scientometrics: An assessment methodology for developing countries. *Scientometrics 7*(3-6), 165–176.

Nicolas Ruffin, N. R. and U. Schreiterer (2017). Case study. science and technology agreements¨ in the toolbox of science diplomacy: Effective instruments or insignificant add-ons? *ELCSID Working Paper, No.6*.

Pei, S., F. Morone, and H. A. Makse (2018). Theories for influencer identification in complex networks. *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, 125–148.

Peixoto, T. P. (2017). Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E 95*(1), 012317.

Prebisch, R. (1962). The economic development of Latin America and its principal problems. *Economic Bulletin for Latin America*.

Priem, J., H. Piwowar, and R. Orr (2022). OpenAlex: A fully open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Saha, S., S. Saint, and D. A. Christakis (2003). Impact factor: A valid measure of journal quality? *Journal of the Medical Library Association 91*(1), 42.

Schott, T. (1988a). International influence in science: Beyond center and periphery. *Social science research 17*(3), 219–238.

Schott, T. (1988b, September). International influence in science: Beyond center and periphery. *Social Science Research 17*(3), 219–238.

Schott, T. (1998, August). Ties between center and periphery in the scientific world-system: Accumulation of rewards, dominance and self-reliance in the center. *Journal of World-Systems Research*, 112–144.

Scott, P. (2015). Dynamics of academic mobility: Hegemonic internationalisation or fluid globalisation. *European Review 23*(S1), S55–S69.

Seth, S. (2009, December). Putting knowledge in its place: Science, colonialism, and the postcolonial. *Postcolonial Studies 12*(4), 373–388.

Shang, J., J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han (2018). Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering 30*(10), 1825–1837.

Shils, E. (1975). *Center and Periphery: Essays in Macrosociology*. University of Chicago Press.

Stringer, M. J., M. Sales-Pardo, and L. A. N. Amaral (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS One 3*(2), e1683.

Sumathipala, A., S. Siribaddana, and V. Patel (2004, December). Under-representation of developing countries in the research literature: Ethical issues arising from a survey of five leading medical journals. *BMC Medical Ethics 5*(1), 5.

Sun, X. and W. Xu (2014, Nov). Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters 21*(11), 1389–1393.

Tickner, A. B. (2013, sep). Core, periphery and (neo)imperialist international relations. *European Journal of International Relations 19*(3), 627–646.

Urbinati, A., E. Galimberti, and G. Ruffo (2021). Measuring scientific brain drain with hubs and authorities: A dual perspective. *Online Social Networks and Media 26*, 100176.

Wagner, C. S. and K. Jonkers (2017, October). Open countries have strong science. *Nature 550*(7674), 32–33.

Wagner, C. S., H. W. Park, and L. Leydesdorff (2015, July). The continuing growth of global cooperation networks in research: A conundrum for national governments. *PLoS One 10*(7), e0131816.

Wuestman, M. L., J. Hoekman, and K. Frenken (2019). The geography of scientific citations. *Research Policy 48*(7), 1771–1780.

Zelnio, R. (2012, May). Identifying the global core-periphery structure of science. *Scientometrics 91*(2), 601–615.

Zhang, L., Z. Cao, Y. Shang, G. Sivertsen, and Y. Huang (2024). Missing institutions in openalex: possible reasons, implications, and solutions. *Scientometrics*, 1–23.

Zhou, P. and L. Leydesdorff (2006). The emergence of china as a leading nation in science. *Research policy 35*(1), 83–104.

Supplemental Information

S1 Top Journals

Assessing the quality of a publication by its venue, often the journal in which it is published is a common practice in academic research (Saha et al., 2003; McKiernan et al., 2019). This approach is based on the premise that the reputation and rigour of the peer-review process of a journal are indicative of the quality of the articles it publishes. Top journals are traditionally identified by their *impact factor*, the average number of citations to publications in that journal over a 2-year window, which is susceptible to temporal and disciplinary variations (Garfield, 2006; Althouse et al., 2009). To control for these, we focus on all publications in each journal and field (OpenAlex Level 1) and find the number of citations to the publication over 5 years (c_5). We then leverage that the journal-specific citation distributions are log-normal (Stringer et al., 2008), and rank each journal in each field and each year by the mean log number of citations over 5 years. Finally, we take the top 50 journals in each field and each year, giving the set of top journals.

Using the yearly top journal set, we identify the fraction of publications from each country in the top journals, and normalize by the overall fraction of the global publications in these top journals (this quantity was decreasing over the time period considered).

S2 Identifying Scientific Ideas

To identify scientific ideas, we follow the methodology introduced in Cheng et al. (2023) (Cheng et al., 2023). We begin by pre-processing OpenAlex texts in several ways. First, we generate our input corpus by combining the abstract and title of each OpenAlex article. Then we remove the last sentence of an abstract if it contains copyright information. Next, we lowercase the text, remove digits, and replace punctuation except commas and periods with spaces. Finally, we use Porter lemmatization on the corpus for all words longer than five characters to collapse different variations of the same word (e.g., singular versus plural forms).

We then identify all publications that mention at least one of the ideas from the master list of 46,535 scientific ideas derived by Cheng et al. using a data-driven phrase segmentation algorithm, AutoPhrase (Shang et al., 2018). This results in a corpus of 1,191,364 publications from 221 countries. Next, we post-process these ideas, removing cases that were first mentioned before 2000 and focusing only on those ideas that were mentioned by only one country in their first year of usage, resulting in 7,327 unique ideas mentioned in 202,932 publications. Finally, we derive a dyadic variable denoting the fraction of ideas whose first usage was in the Origin country and then were later used in the Destination Country sometime between 2000 and 2022.

S3 International citation preference network

Table S1. Fixed-effect multinomial logit regression for 1990-2017. Model coefficients labelled by *p*-value. Standard errors in parentheses.

	Dependent variable: Citation preference					
	Model					
	(1)	(2)	(3)	(4)	(5)	(6)
Citation Preference : Positive						
Intercept	−15.92***	−16.64***	−17.22***	−17.4***	−13.55***	−15.57***
	(−16.61,−15.23)	(−17.34,−15.93)	(−17.99,−16.44)	(−18.18,−16.62)	(−14.34,−12.76)	(−16.4,−14.73)
	S.E. 0.35; p-v 0.0	S.E. 0.36; p-v 0.0	S.E. 0.4; p-v 0.0	S.E. 0.4; p-v 0.0	S.E. 0.4; p-v 0.0	S.E. 0.43; p-v 0.0
Log origin GDP per capita	2.27***	2.05***	1.68***	1.71***	−0.56***	−0.34***
	(2.2,2.33)	(1.98,2.12)	(1.61,1.75)	(1.64,1.78)	(−0.67,−0.45)	(−0.46,−0.22)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.06; p-v 0.0	S.E. 0.06; p-v 0.0
Log target GDP per capita	2.07***	1.89***	1.46***	1.5***	−0.79***	−1.42***
	(2.01,2.13)	(1.82,1.95)	(1.39,1.53)	(1.43,1.57)	(−0.9,−0.69)	(−1.53,−1.31)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.05; p-v 0.0	S.E. 0.06; p-v 0.0
Log origin population	2.17***	2.36***	2.19***	2.26***	0.14*	0.27***
	(2.11,2.23)	(2.29,2.42)	(2.12,2.26)	(2.19,2.33)	(0.03,0.24)	(0.16,0.38)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.05; p-v 0.0105	S.E. 0.06; p-v 0.0
Log target population	1.97***	2.14***	1.91***	1.97***	−0.22***	−0.06
	(1.91,2.02)	(2.08,2.2)	(1.84,1.97)	(1.9,2.04)	(−0.32,−0.11)	(−0.17,0.05)
	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.05; p-v 0.0001	S.E. 0.06; p-v 0.3037
Physical distance		−0.79***	−0.56***	−0.55***	−0.03	−0.03
		(−0.84,−0.75)	(−0.6,−0.51)	(−0.6,−0.51)	(−0.08,0.03)	(−0.09,0.02)
		S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.03; p-v 0.3344	S.E. 0.03; p-v 0.2762
Same continent		0.07	0.11	0.14*	0.32***	0.42***
		(−0.04,0.18)	(−0.01,0.23)	(0.02,0.26)	(0.18,0.46)	(0.27,0.57)
		S.E. 0.06; p-v 0.2189	S.E. 0.06; p-v 0.0769	S.E. 0.06; p-v 0.0264	S.E. 0.07; p-v 0.0	S.E. 0.08; p-v 0.0
Same official language			1.25***	1.2***	0.61***	0.53***
			(1.14,1.35)	(1.1,1.31)	(0.49,0.72)	(0.41,0.65)
			S.E. 0.05; p-v 0.0	S.E. 0.05; p-v 0.0	S.E. 0.06; p-v 0.0	S.E. 0.06; p-v 0.0
Field similarity			3.19***	3.21***	1.63***	2.14***
			(2.99,3.39)	(3.01,3.41)	(1.43,1.82)	(1.93,2.35)
			S.E. 0.1; p-v 0.0	S.E. 0.1; p-v 0.0	S.E. 0.1; p-v 0.0	S.E. 0.11; p-v 0.0
Bilateral research agreements				−0.1***	−0.18***	−0.17***
				(−0.12,−0.08)	(−0.2,−0.16)	(−0.19,−0.15)
				S.E. 0.01; p-v 0.0	S.E. 0.01; p-v 0.0	S.E. 0.01; p-v 0.0
Log collaboration strength					2.73***	2.57***
					(2.62,2.85)	(2.45,2.69)
					S.E. 0.06; p-v 0.0	S.E. 0.06; p-v 0.0
Origin top journal fraction						−0.29***
						(−0.4,−0.18)
						S.E. 0.06; p-v 0.0
Target top journal fraction						1.51***
						(1.39,1.63)
						S.E. 0.06; p-v 0.0
Citation Preference : Negative						
Intercept	−14.79***	−14.49***	−11.41***	−11.38***	−10.5***	−10.63***
	(−15.09,−14.5)	(−14.79,−14.2)	(−11.72,−11.11)	(−11.69,−11.07)	(−10.81,−10.18)	(−10.96,−10.3)
	S.E. 0.15; p-v 0.0	S.E. 0.15; p-v 0.0	S.E. 0.16; p-v 0.0	S.E. 0.16; p-v 0.0	S.E. 0.16; p-v 0.0	S.E. 0.17; p-v 0.0
Log origin GDP per capita	2.64***	2.56***	2.48***	2.48***	2.1***	1.43***
	(2.6,2.69)	(2.52,2.61)	(2.43,2.53)	(2.43,2.53)	(2.04,2.16)	(1.37,1.5)
	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0
Log target GDP per capita	1.14***	1.13***	0.95***	0.94***	0.57***	0.89***
	(1.11,1.17)	(1.1,1.16)	(0.92,0.99)	(0.91,0.98)	(0.52,0.63)	(0.84,0.95)
	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0

Log origin population	2.66***	2.6***	2.53***	2.51***	2.13***	2.17***
	(2.62,2.7)	(2.56,2.64)	(2.48,2.57)	(2.47,2.55)	(2.08,2.19)	(2.11,2.23)
	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0
Log target population	2.44***	2.41***	2.25***	2.23***	1.86***	1.79***
	(2.4,2.48)	(2.37,2.45)	(2.21,2.29)	(2.19,2.28)	(1.8,1.91)	(1.73,1.85)
	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0
Physical distance		−0.19***	−0.07***	−0.07***	0.03	0.08***
		(−0.22,−0.15)	(−0.11,−0.04)	(−0.11,−0.04)	(−0.01,0.06)	(0.05,0.12)
		S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.1488	S.E. 0.02; p-v 0.0
Same continent		−0.72***	−0.71***	−0.7***	−0.7***	−0.69***
		(−0.8,−0.64)	(−0.79,−0.63)	(−0.79,−0.62)	(−0.78,−0.61)	(−0.78,−0.6)
		S.E. 0.04; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.04; p-v 0.0	S.E. 0.04; p-v 0.0
Same official language			−0.63***	−0.61***	−0.75***	−0.74***
			(−0.73,−0.53)	(−0.71,−0.51)	(−0.85,−0.65)	(−0.84,−0.63)
			S.E. 0.05; p-v 0.0	S.E. 0.05; p-v 0.0	S.E. 0.05; p-v 0.0	S.E. 0.05; p-v 0.0
Field similarity			0.73***	0.73***	0.53***	0.62***
			(0.67,0.79)	(0.67,0.79)	(0.47,0.6)	(0.56,0.69)
			S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0	S.E. 0.03; p-v 0.0
Bilateral research agreements				0.01*	−0.01	−0.01
				(0.0,0.02)	(−0.02,0.0)	(−0.02,0.0)
				S.E. 0.01; p-v 0.0488	S.E. 0.01; p-v 0.0708	S.E. 0.01; p-v 0.1369
Log collaboration strength					0.4***	0.48***
					(0.37,0.44)	(0.44,0.52)
					S.E. 0.02; p-v 0.0	S.E. 0.02; p-v 0.0
Origin top journal fraction						0.88***
						(0.82,0.94)
						S.E. 0.03; p-v 0.0
Target top journal fraction						−0.75***
						(−0.8,−0.69)
						S.E. 0.03; p-v 0.0
Note:		*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$				
Observations	930798	844882	744760	744760	744760	744760
Pseudo R^2	0.5613	0.5793	0.5954	0.5963	0.6198	0.6386
Log Likelihood	-32842.41	-30453.84	-28640.06	-28580.11	-26918.1	-25582.81
LLR χ^2	84024.71*** (d.f.=62.0)	83875.63*** (d.f.=66.0)	84307.51*** (d.f.=70.0)	84427.42*** (d.f.=72.0)	87751.43*** (d.f.=74.0)	90422.0*** (d.f.=78.0)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

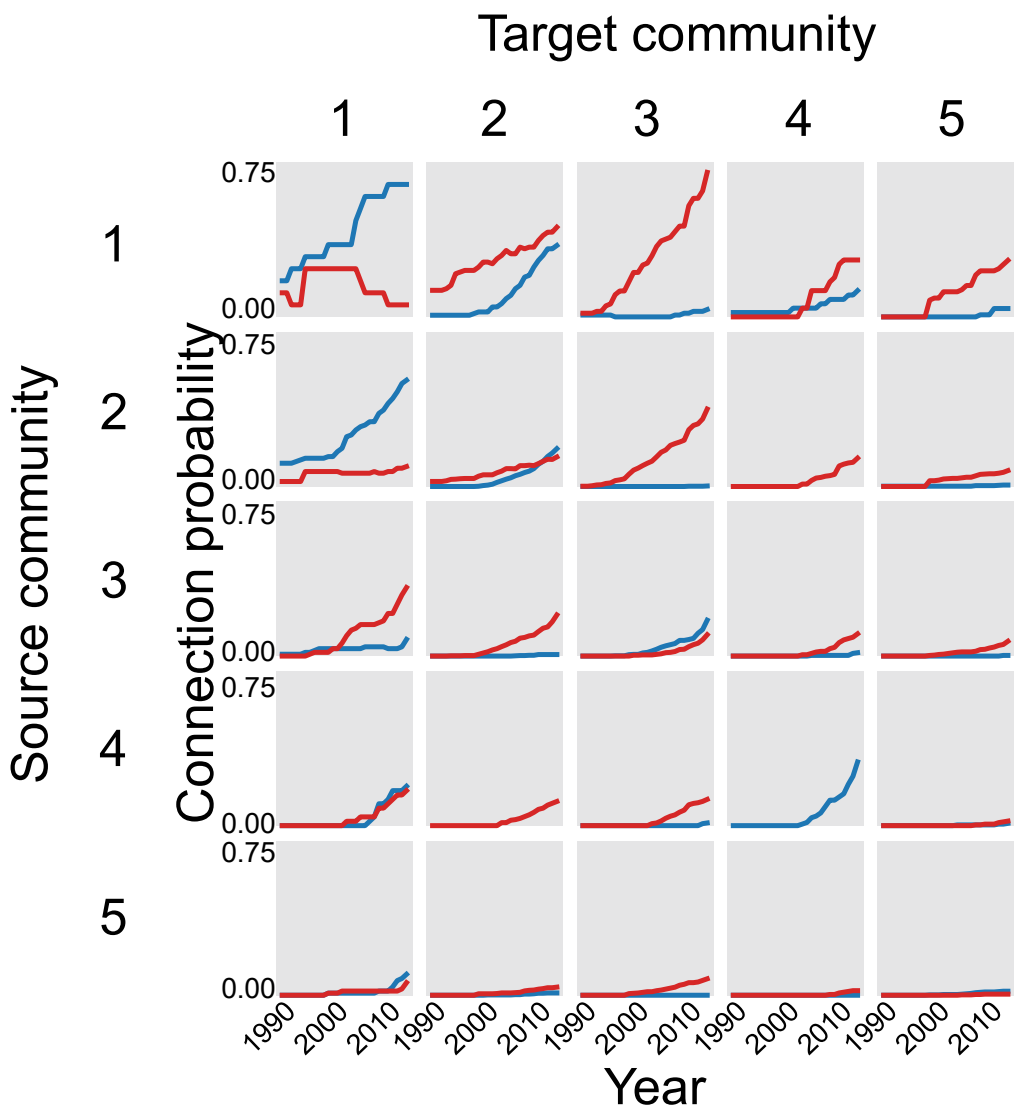


Figure S1. International network fragmentation. The probability of a positive (blue) or negative (red) directed edge from a country in the source community (rows) to a country in the target community (columns) from 1990 until 2017 (x-axis).

https://doi.org/10.51408/issi2025_161

The Interaction between Scientific Research and Policy in The Field of Supply Chain: An Empirical Analysis Based on Overton Data

Li Jiangbo[1], Li Jiake[2], Mu Yingyu[3], Li Jian[4]

[1]*jiangbosky@sina.com*
Business College, Qingdao University, Qingdao 266100 (China)
Research Institute for Science and Technology System and Institutional Innovation of Qingdao, Qingdao 266100 (China)

[2] *lijiake@qdu.edu.cn,* [3] *muyingyu@qdu.edu.cn,* [4]*lj13181433265@163.com*
Business College, Qingdao University, Qingdao 266100 (China)

Abstract

Against the backdrop of accelerating global economic integration and digital transformation, the complexity of supply chain management has continuously escalated, with a significant increase in policy dependency. This necessitates a systematic investigation into the interaction between academic research and policy-making to enhance the scientific rigor and effectiveness of decision-making.This study integrates data from the Overton policy database (covering policy documents from 1991 to 2025) and the Web of Science (WOS) academic database (including research articles from 1978 to 2024) by matching Digital Object Identifiers (DOIs) and policy IDs. A total of 116,193 supply chain-related academic papers (including 4,379 papers cited by policy documents) and 237,849 policy documents (including 8,556 documents citing academic papers) were identified. Empirical analysis was conducted using the Mann-Whitney U test and Spearman correlation analysis.The findings reveal that academic papers cited by policy documents (n = 4,379) had an average citation count of 110.8 in the WOS Core Collection, which is significantly higher than the average citation count of 29.8 for non-cited papers (n = 111,814), representing a 3.7-fold difference. Similarly, policy documents citing academic papers (n = 8,556) had an average citation count of 8.7 in the policy domain, which is 4.3 times higher than that of non-citing documents (n = 229,293).Correlation analysis indicates a weak positive association between academic citation impact and policy citation frequency suggesting that policy documents tend to reference research with high immediate relevance, whereas academic influence requires long-term accumulation.The study underscores a bidirectional synergy between academia and policy-making in the supply chain domain: policy documents enhance their scientific validity and authority by citing high-impact academic research, while policy needs drive academic research toward practical issues. This study quantitatively assesses the reciprocal citation relationship between science and policy in the supply chain field, providing empirical evidence for the policy translation of academic research findings.

Introduction

In modern enterprise management, as global economic integration and market competition intensify, enterprises no longer compete independently but as part of a supply chain comprising multiple businesses and relationship networks. The American Supply Chain Management Association (APICS/SCC) defines the supply chain as a value-added business network centered on a core enterprise, encompassing material acquisition, processing, and product delivery. It operates through the control of information, logistics, and capital flows, forming a logistics chain, information chain, and capital chain.

In globalization, supply chain management (SCM) has become a key academic focus. The supply chain revolves around a core enterprise, controlling information, logistics, and capital flows from raw material procurement to product manufacturing and final delivery through a sales network. It forms a functional network linking suppliers, manufacturers, distributors, retailers, and consumers. Emphasizing cross-organizational and cross-regional resource coordination, it optimizes logistics, information, and capital flows to reduce costs, enhance efficiency, and mitigate risks. Since the 1990s, supply chain research has advanced in theory and practice, expanding into areas like collaboration, finance, risk management, and sustainability, bringing significant economic and social benefits. With the rise of information technology and big data, it has integrated multidisciplinary foundations, including management science, economics, and sociology, while leveraging emerging technologies such as blockchain, IoT, AI, and cloud computing to enhance flexibility, intelligence, and transparency. Recent global crises, including pandemics, geopolitical conflicts, and natural disasters, have exposed supply chain vulnerabilities, driving research on resilience and sustainability. Scholars explore risk identification, early warning, and response strategies to mitigate disruptions and balance public interests with corporate profits (Chowdhury et al., 2021). As a result, supply chain research now extends beyond operations and costs to encompass environmental sustainability, social responsibility, resilience, and risk management.

In summary, supply chain research is vital for enterprise management, global economic efficiency, and sustainable development. As economic globalization and digital transformation accelerate, supply chain operations will grow more complex and increasingly interconnected with macro policies. Policy factors—such as tariffs, trade agreements, industrial support, regulations, and risk management—profoundly impact supply chain stability, efficiency, and sustainability. With environmental and carbon neutrality goals, policies have become key external influences. Green supply chain theory highlights how regulations (e.g., carbon emission controls, environmental standards) shape corporate sustainability (Ji et al., 2024b). This policy-driven pressure reshapes business models, driving new supply chain strategies

that integrate social responsibility. Effective policies optimize resource allocation, foster sustainability, enhance resilience, and promote fair competition and social welfare.

To navigate the evolving global supply chain landscape, academia and policymakers must strengthen interdisciplinary and technological collaboration to develop a sustainable, inclusive policy system. Analyzing the supply chain-policy interaction can yield innovative frameworks and tools to enhance efficiency, achieve sustainability goals, and address future uncertainties, fostering global economic prosperity and social welfare. For researchers, systematically examining policy impacts on supply chains is crucial for informing scientific policy-making and optimizing corporate strategies. The link between policy and science is key: policies cite high-quality research to guide institutions and allocate resources, while academic findings support supply chain optimization and transformation. This synergy enhances efficiency, resilience, and sustainability at both technical and institutional levels, driving balanced economic and social development.

In recent years, policymakers have placed increasing emphasis on the use of research evidence in policymaking (Hui et al.,2020; Obuku et al., 2018). At the same time, in academia, researchers are thinking about how to conduct research in such a way as to better provide evidence for policy-making. Amid accelerating globalization and digitalization, exploring the policy-science connection has become crucial for advancing supply chain research. However, previous studies faced challenges due to the lack of a reliable global data source for analyzing this relationship. In 2019, the Overton policy document database was introduced, compiling policy documents and their citations of academic papers. This study leverages Overton, which includes records from government agencies, think tanks, and intergovernmental organizations, to examine the interaction between scientific research and policy in the supply chain field. The influence of academic findings may be reflected in policy document citations.

literature review

Overview of studies on the connection between science and policy

In recent years, the phenomenon of cross-domain knowledge diffusion from science to policy has become increasingly evident (Nay & Barré-Sinoussi, 2022). This refers to the process of introducing scientific research results into policy formulation and implementation to solve specific problems and challenges (Hodges et al., 2022). In this process, scientific research results need to be translated into specific policies and practical measures to meet the needs of policymakers and implementers (Watson, 2005). Research institutions (e.g., universities), as well as researchers, are working

to ensure that their research is considered in the policy-making process (Ray et al., 2021). However, in previous studies, the disconnect between science and policy is a long-standing problem, in which policymakers may miss important scientific insights and erroneous scientific advice may affect decision-making.

Yin et al. (2021) pointed out that the reason for the limited systematic understanding of the connection between science and policy is the lack of reliable data worldwide, making it difficult to reliably track the co-evolution of policy and science on a global scale. As a result, there was relatively little early research on the science-policy interface. For example, Haunschild et al. (2016) explored the feasibility of policy documents as a source for measuring the social impact of scientific research by examining the frequency of references to climate change-related scientific research in policy-related documents. Using data from Altmetric.com, Haunschild and Bornmann (2017) investigated the extent to which articles indexed by the Web of Science (WOS) are mentioned in policy documents. They found that less than 0.5% of articles are mentioned at least once in relevant policy documents. Vilkins and Grant (2017) conducted a study using documents from policy-focused Australian government departments. They found that the majority of citations were peer-reviewed journal articles, federal government reports, and Australian business information. The study also suggests that 'the chances of being cited may increase if the academic research is open access.' Additionally, Newson et al. (2018) explored the current status of research citations in policy documents on childhood obesity in New South Wales, Australia, and its feasibility as an indicator of research impact by analyzing policy documents from 2000 to 2015, revealing how scientific research is adopted by policy and its practical impact on policy development.

But in 2019, the new OVERTON policy document database was released, which includes links to research papers cited in policy documents (Overton,2020). Yang et al. (2020) define policy documents in this context as carriers of policy. The OVERTON database provides a channel for policy science researchers to study the main content of policies, policy-making processes, and policy tools.Policy documents are an important data source to investigate the social impact of research (Drongstrup et al., 2020; Yu et al., 2020).Since then, research on the science-policy nexus has gradually increased.Drongstrup et al. (2020) found that economics articles published in high-level journals were more likely to be cited in policy documents than those published in low-level journals.Yin and Gao used Overton data to analyze the connection between science and policy regarding COVID-19.They found that "many policy documents on the COVID-19 pandemic substantially cite the latest, peer-reviewed, high-impact science. Policy documents that cite science are particularly highly cited in the policy field. At the same time, there are differences in the use of science by different decision-making bodies. The tendency of policy

documents to cite science seems to be mainly concentrated in intergovernmental organizations (IGOs) such as the World Health Organization (WHO), but very few in national governments, because they mainly cite science indirectly through IGOs.Cheng et al. (2021) studied the co-evolutionary relationship between scientific research and policy making in China during the early stages of the COVID-19 epidemic, and proposed a science-policy coevolution model (CEM) to explain the dynamic interaction in public health emergencies. Bornmann et al. (2022) discussed the question of how climate change research is connected to policy. They pointed out that intergovernmental organizations and think tanks pay more attention to climate change and have issued more climate change policy documents than expected. The authors found that climate change papers cited in climate change policy documents were cited much more often on average than climate change papers not cited in these documents. Both scientific papers and policy documents focus on similar areas of climate change research: biology, earth sciences, engineering, and disease science.

In addition to this, there are other studies that examine the relationship between science and policy from different perspectives. Fang et al. (2020) focused on hot research topics reflected in papers cited in policy documents. Brandts-Longtin et al. (2022) explored the potential impact of predatory journal articles on policy and guidance documents, analyzed how these low-quality scientific studies infiltrated policy areas through a cross-sectional study design, and evaluated their possible consequences for public decision-making. Cristofoletti, Evandro Coggo, et al. (2023) revealed the interactive relationship between scientific research and policy making by analyzing the citations of research related to projects funded by the Sao Paulo Research Foundation (FAPESP) in policy documents, and proposed a new methodological framework to evaluate the policy impact of research. Yoshida et al. (2024) explored the importance of gray literature in the scientific policy process and applied research, especially in supplementing the evidence and knowledge of peer-reviewed literature. Llewellyn et al. (2023) explored the translation path of scientific research results in health policy, and proposed an evaluation framework that links translational research publications with policy literature through innovative bibliometric methods. Van Elsland et al. (2024) analyzed the policy impact of the research of the Imperial College COVID-19 Response Team (ICCRT) during the epidemic, and explored how its research results influenced global and British policy decisions through different dissemination channels. Ma and Cheng (2024) describe the citation of Public Administration and Policy (PAP) academic papers within policy documents and find that the three dimensions of collaborative teams, interdisciplinary interactions, and disruptive paradigms are all influential factors that increase the citation rate of academic papers in this field within policy documents,

but the relationship between them is not linear. Using publication data on COVID-19 topics, Hu et al.(2024b) found a positive correlation between the interdisciplinarity of scientific publications and the attention given to them in policy documents in almost all fields.

Overview of studies on Overton

In previous studies, data on policy documents and policy citations could only be obtained from databases of companies such as Altmetric and PlumX. In 2019, the Overton database emerged to change this situation, aiming to become the largest policy document and citation database. In the OVERTON database, policy documents are defined as "documents written very broadly primarily for or by policymakers". Overton includes documents from governments, think tanks (i.e. research institutions that conduct research and advocacy on climate change), non-governmental organizations (NGOs), and intergovernmental organizations (IGOs, i.e. organizations composed of countries). The database includes not only various bibliographic information of policy documents (such as titles and appearances), but also citation links between policies and science and between policy documents in the database itself. The Overton database uses text mining methods to identify citation relationships.

Yin and Gao studied the reliability of science policy citations in the Overton database by comparing them with the citation links provided by the Microsoft Academic Graph database. The results showed that "although the two datasets were collected for different purposes using different methods and techniques, independent measurements on the two datasets showed significant consistency."

Since then, there has been a gradual increase in the number of studies based on overton databases. Cabral and Salles-Filho (2024) analyzed the evolution of global artificial intelligence (AI) policy documents and their scientific basis through the Overton policy document database. The study found that the number of AI policy documents has increased significantly since 2018, and the United States, the European Union and international organizations have played a leading role in policy making. Fourough Rahimi, F., & Danesh, F. (2024) conducted a scientometric analysis of 2,493 political documents related to open government data in the Overton database from 2007 to 2023 based on scientometric indicators and content analysis. The study found that the Organization for Economic Cooperation and Development (OECD) and the Guardian News Agency performed outstandingly in terms of the number of citations, and there was a significant positive correlation between GDP and the number of open government data policy documents at the national level. Haunschild, R et al. (2023, April) used the OVERTON database to explore the extent to which public policy and administrative research has influenced policy

departments. By analyzing the citations of public policy and administrative research in policy documents, it was revealed which research contributed the most to policy reports and decisions, and which policy institutions used research literature more frequently to support their policy decisions. Szomszor and Adie (2022) explored the citation of academic literature in policy making by analyzing the Overton policy document database. Ren and Yang (2023) used the OVERTON database to explore the characteristics of the diffusion of scientific knowledge into the policy field and found that the intensity and breadth of the diffusion conformed to the power law distribution, while the diffusion speed conformed to the log-normal distribution. Huang et al. (2022) used data from the Overton database to study the association between scientific collaboration and its policy impact in the field of library and information science (LIS). Through quantitative analysis of policy citations in LIS research, the important role of international collaboration in enhancing the impact of research policies was revealed. Xu and Zong (2023) used overton data to test the effect of international research cooperation on policy impact through PSM method, and the results of the study showed that international research cooperation has a significant positive effect on the policy impact of scientific research.

Other scholars have studied overton by combining it with other data sources. Dorta-González et al. (2024b) used Altmetrics and the Overton database to explore how scientific research results affect policy making and analyzed the citations of nearly 125,000 articles from 434 public policy journals. The study found that news and blog mentions, social media participation, and open access publications can significantly increase the likelihood of research articles being cited in policy documents, while non- open source articles have a lower chance of being cited in policies. Pinheiro et al. (2021b) used publication data from the Framework Programs for Research and Technological Development (FPs) to investigate the relationship between interdisciplinarity at the paper level and policy impact measured by policy citation data from the Overton database. The results show that measuring the use of policy-related literature based on the OVERTON database can benefit research. The OVERTON database can capture the interaction between science and policy and the contribution of these interactions to the larger decision-making process. Jonker and Vanlee (2024) reveal for the first time the media mentions and policy citations of all active scholars at Dutch-speaking universities in Belgium by linking data from FRIS, BelgaPress and Overton.

In summary, the Overton platform brings together a large number of academic documents, policy documents, patents, etc., and can track and analyze the citations and application backgrounds of scientific research results in the policy field. The database provides a valuable analytical tool for the interaction between research papers and policy documents, especially in measuring the impact of academic

research on policy making. For researchers, it can help them understand how their research has aroused public attention and ultimately turned into policy actions; for policymakers, it provides a reference channel to help them formulate more scientific and effective policies based on the latest academic research. Through Overton, researchers and policymakers can clearly see how academic results affect policy documents and actual decisions. In general, the Overton database has played a positive role in promoting interaction and communication between academic research and policy and enhancing the social influence of scientific research results.

Research Objectives and Research Questions

Given the increasing need for science-policy interaction in the supply chain domain and the existing research gaps, this study aims to systematically examine the relationship between academic research and policy-making. By integrating empirical data from the Overton policy database and the Web of Science (WOS) academic database, this study pursues three key objectives. First, it seeks to quantify policy citation preferences by investigating whether supply chain policy documents tend to cite high-impact academic papers and assessing the difference in academic influence between cited and non-cited papers. Second, it examines the correlation between policy influence and the citation of academic research, analyzing whether policy documents referencing academic studies receive greater recognition within the policy domain. Lastly, the study explores the broader interaction between science and policy by identifying statistical associations between academic citations and policy citations, thereby evaluating the extent to which academic influence affects policy adoption.

To achieve these objectives, this study addresses the following research questions:

RQ1: Do academic papers cited in supply chain policy documents exhibit significantly higher academic influence (e.g., citation counts) than those that are not cited?

RQ2: Do supply chain policy documents that reference academic papers have greater policy influence (e.g., citation frequency by other policy documents) than those that do not?

RQ3: Is there a significant correlation between the academic influence of a paper and its likelihood of being cited in policy documents?

Data Acquisition

Data Source

In order to explore the science and policy in the field of supply chain, in this study, we chose the Overton database as our data source for obtaining policy documents

and the academic papers they cited. Overton defines policy documents as "studies, briefs, reviews, or reports " written with the purpose of influencing or changing policy, and provides scientific and policy citations in each document. For each policy document, the Overton database has a unique policy ID code to match it. The Overton database contains links to academic papers through digital object identifiers (DOIs), and "academic" papers in Overton have a unique DOI. As for the source of academic papers related to the supply chain field, we chose the Web of Science academic paper database as the source of academic papers.

Data processing

In order to find the most relevant scientific research results for supply chain policy, we searched for relevant academic research papers using the keyword "Supply Chain" in the " Search Academic Papers" search window in the Overton database. According to our search results on January 10, 2025, a total of 6,442 relevant academic papers were obtained, with publication dates ranging from 1978 to 2024. In our subsequent research, we used these 6,442 academic papers as a paper subset to represent all academic papers in the field of supply chain that have been cited in policy documents.

After obtaining 6442 academic papers related to supply chain policies, we then used the Overton database to obtain the policy collection that cited these 6442 academic papers in the policy library. As of January 10, 2025, a total of 12692 policy documents that cited the above academic papers were obtained. Since our main focus is policy documents, we follow Overton's advice and further filter the file type, using only "publications" (accounting for 90.5% of the total number of documents) and removing other types such as "working papers". Finally, 11485 policy documents that cited these academic papers were obtained. Subsequently, the data were analyzed and processed by a computer program, which detected and removed duplicate records from the data, removing a total of 554 duplicates, thereby effectively reducing data noise. As a result of the above processing, 10,931 policy documents in the field of supply chain with the document type of “Publication” were obtained. For each policy document, we have its title, original URL, publication date, document type, policy source and subject classification, as well as a unique policy ID code, the number of times it was cited by other policies (including the average number of policy citations after removing the citations from the policy source agency itself and the average number of policy citations without removing the citations from the policy source agency itself). The distribution of publication years and source types of these 10,931 policy documents are shown in Figures 1, 2 respectively.

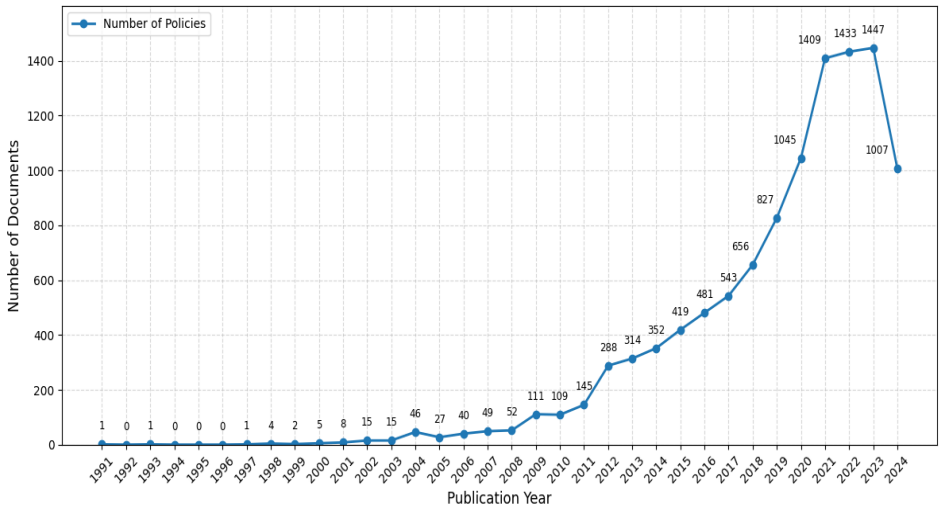


Figure 1. Distribution of the release years of the 10,931 policy documents.

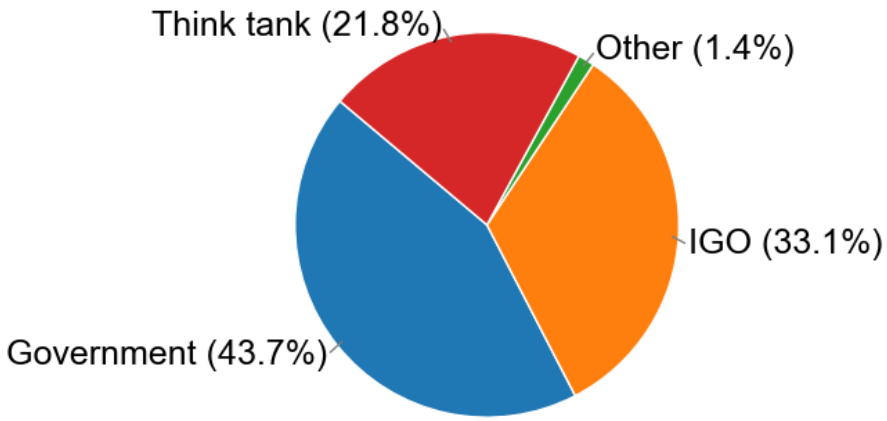


Figure 2. Distribution of source types of 10,931 policy documents.

In order to obtain more policy documents related to the supply chain field (regardless of whether they cite academic papers), we searched for documents using the exact phrase "Supply Chain" in the "Search Policy Documents" window in the Overton database. Similar to the above, we only selected policy documents with the file type "Publication". As of our search time on January 10, 2025, we retrieved a total of 264,759 relevant policy documents. In order to be consistent with the publication time of the previous 10,931 policy documents, we again limited the time and only retained the supply chain field policy documents with the type of "Publication" published from 1991 to 2025. Subsequently, the data were processed by a computer program to detect and remove duplicates. This process resulted in the deletion of a

total of 12,972 duplicate records, thereby effectively removing data noise. Consequently, 237,849 policy documents in the domain of supply chain management with the document type of "Publication" were obtained. Similarly, for each policy document, we have its title, original URL, publication date, document type, policy source and subject classification, as well as unique policy ID code, number of citations by other policies, and other information. The distribution of the year of release, and the source type of these 237,849 policy documents are shown in Figures 3, 4, respectively.

This study finds that both sets of supply chain policy releases show a similar evolutionary trajectory in the time dimension: firstly, the number of annual policies remains very low from the early 1990s to around 2005; then, from around 2006 onwards, there is a gentle rise in the data and an accelerated climb after 2010, signalling a growing interest in supply chain issues. Between 2015 and 2020, both sets of data show rapid growth and reach relative peaks around 2020, respectively, suggesting a concentrated burst of policy interest during this period. After reaching their peaks, the number of releases dropped off in 2023 and 2024, although they are still well above the levels of the earlier years. This downward trend may be related to factors such as data not yet being fully collected, a change in policy focus, or the period of concentrated policy releases having passed.

Overall, the chart reflects the explosive growth of policies in the supply chain sector from few to many over the last decade or so, with a peak followed by a phase of relative decline but still a high base. Meanwhile, the comparison of the pie charts shows that the distribution of source types has changed somewhat as the size of the data has increased, with government sources accounting for a relatively higher proportion of the second set of data, and other types (igo, think tank, etc.) accounting for a relatively lower proportion. This may be due to the fact that policy documents from the government are less likely to be cited in academic papers. Previous scholars have come to similar conclusions. (Yin et al., 2021b) After our inspection, 8,556 of the 10,921 policy documents initially obtained were included in these 237,849 policy documents. In subsequent research, we decided to use these 8,556 policy documents to represent the set of policy documents in the field of supply chain that cited academic papers.

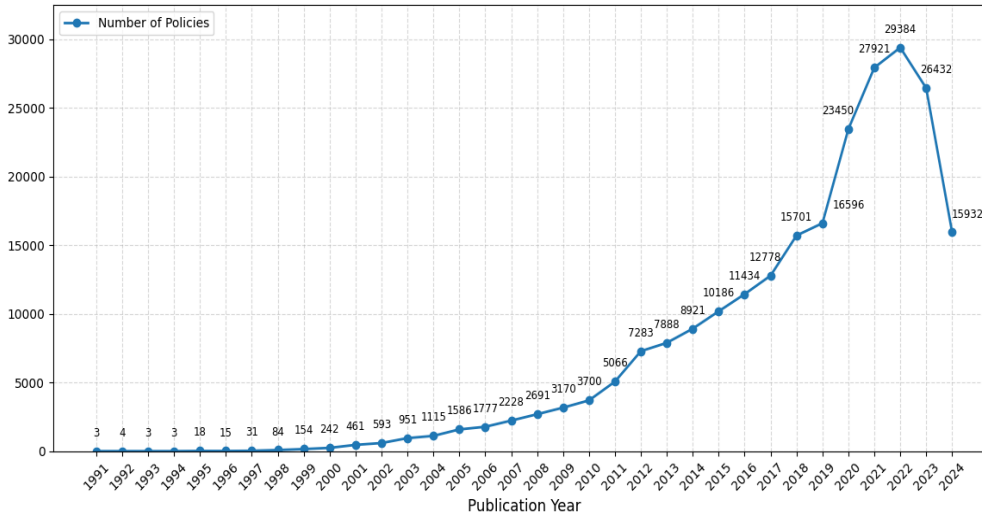


Figure 3. Distribution of the release years of the 237,849 policy documents.

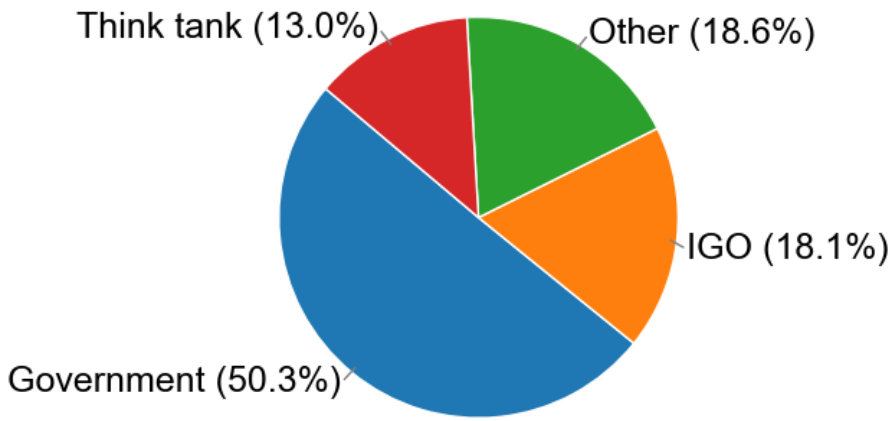


Figure 4. Distribution of source types of 237,849 policy documents.

In order to obtain more academic papers related to the supply chain field (regardless of whether these academic papers have been cited in policy documents), we chose the Web of Science academic paper database as the source of academic papers. We searched all databases of WOS using the keyword "Supply Chain". Since our research object is mainly academic papers, we retained the results of the document types "paper" and "review paper". In order to be consistent with the academic papers obtained from the Overton database above, we limited the publication time of the search results to 1978 to 2024. Finally, 146,558 supply chain-related academic papers were retrieved. After removing 21,273 data without DOI numbers (about 14.5% of the total data) and 9,092 duplicate data (about 7.2% of the total data), we

finally obtained 116,193 academic papers. For each paper, we will obtain information about its title, author list, publication date, file type, DOI number , abstract , and number of citations (including the number of citations in the Web of science core database and the number of citations in all Web of science databases).Match the academic papers cited by Overton with those obtained from the Web of science database through DOI numbers. The matching results show that 4379 of the 6442 academic papers obtained from the Overton database are also included in the Web of science. Therefore, we can obtain the citations of these academic papers by other papers in the Web of science database.

In this study, we use the number of times an academic paper is cited by other papers to measure the quality of an academic paper. The more times a paper is cited, the higher its academic influence, that is, the higher its quality. Similarly, we use the number of times a policy document is cited by other policies in the Overton database to measure the quality of the policy. The more times a policy is cited by other policies, the more influence it has, that is, the higher its quality.

As shown above, in our study, we have two sets of academic papers and two sets of policy documents. We regard 8556 policy documents that exist in both policy sets as policy documents that cite academic papers, and 4379 academic papers that exist in both academic paper sets as academic papers cited by policies. For comparison, we remove the 8556 policy documents from the 237849 policy documents and the remaining 229293 policy documents as policy documents that do not cite academic papers, and remove 4379 academic papers from the 116193 academic papers obtained from Web of science. The 111814 papers represent academic papers that are not cited by policies. In the following research, we hope to measure the quality of the two sets of policies or academic papers by the citations they receive, and ultimately find out the mutual influence of academic papers and policy documents in the field of supply chain.

Results

Correlation analysis between the number of academic citations and the number of policy citations of academic papers

In the study of communication and policy impact, the number of citations of academic papers can be one of the important indicators of their impact. Based on the 4379 academic papers cited by policy obtained above, we analyse the correlation between the number of citations of these papers in the academic field (divided into the number of citations in the core database and the number of citations in all databases, provided by the Web of Science core database) and the number of

citations in the policy field (provided by the Overton database), in order to explore the correlation between the academic influence and the policy influence.

Our dataset contains the following three columns of key fields:1. DOI: the unique identifier of each academic paper, which is used to distinguish different papers;2. the number of times a paper has been cited by core databases: i.e. the number of academic citations, reflecting the influence of the paper in academia;3. the number of citations by policies: i.e. the number of policy citations, reflecting the influence of the papers in policy making.

Preliminary checking of the data shows that there are no missing values in these fields, indicating that the data are complete and can be used directly for analysis. For correlation measures, we use the Spearman Correlation Coefficient as a correlation measure. Spearman Correlation Coefficient is suitable for non-linear or non-normally distributed data, and can measure the monotonic relationship between two variables. By calculating the Spearman Correlation Coefficient between the number of citations in core databases and the number of citations in policies, we can find out the strength of the correlation between the two.

By calculating the correlation coefficients between the number of citations in WOS core database and the number of citations in all WOS databases on the number of citations of academic papers by policy, the following results are obtained: the Spearman's correlation coefficients between the number of citations in two kinds of WOS and the number of citations of academic papers by policy are all 0.25. The two sets of results are shown in Fig. 5 and Fig. 6 in the following figures. This result shows that there is a weak positive correlation between the number of citations in core databases and the number of policy citations, and a weak positive correlation between the number of citations in all databases and the number of policy citations. There is also a weak positive correlation between ‘number of citations in all databases’ and ‘number of citations in policy’, i.e., papers with more academic citations are more likely to be cited in policy documents to a certain extent. The weak correlation may be partly due to differences in citation motivation: academic citations are mainly motivated by research background and theoretical support, while policy citations are more driven by practical needs and social issues. There may be a difference in emphasis between the two. In addition, temporal factors may also play a role: academic citations usually take a long time to accumulate, whereas policy citations may be closely related to unexpected events, leading to differences in the temporal distribution of citation patterns.

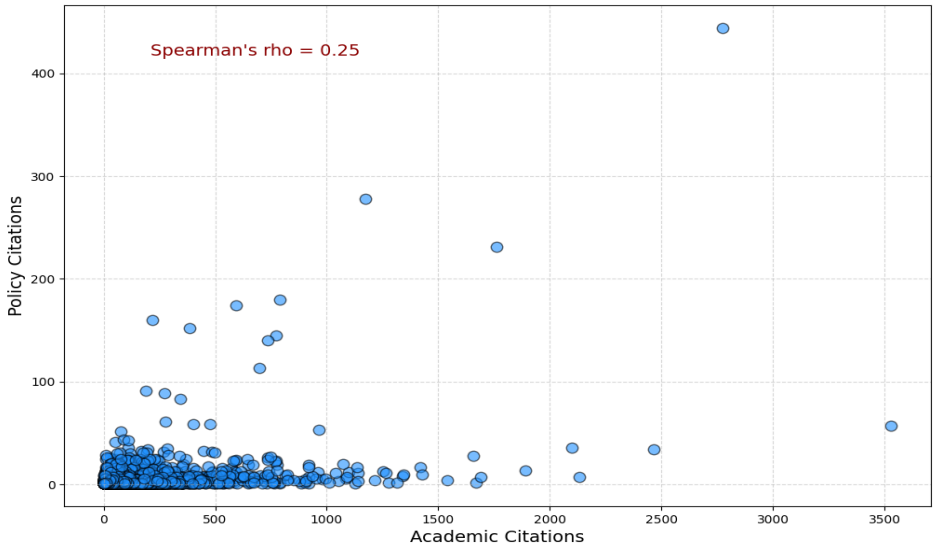


Figure 5. Scatterplot of the correlation analysis between the number of citations of academic papers by the WOS core database and the number of citations by policy.

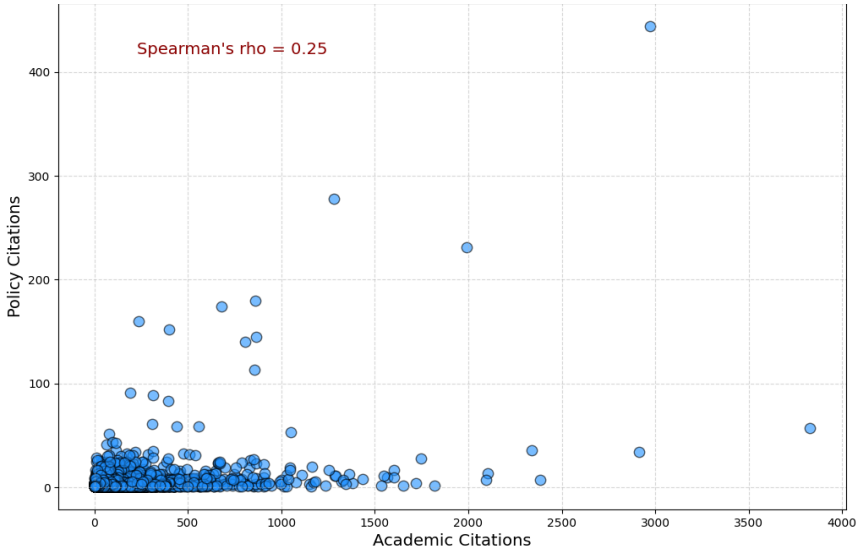


Figure 6. Scatterplot of the correlation analysis between the number of citations of academic papers by all WOS databases and the number of citations by policy.

Difference in the number of citations to other papers between academic papers cited by the policy and those not cited by the policy

Based on the above, two collections of policy documents and a collection of academic papers are compared in terms of the number of citations

respectively.Processing the acquired paper data, it can be obtained that the 4379 academic papers cited by the policy that exist in both paper collections have an average of 110.822 citations by papers in the Web of science core database and 122.574 citations by papers in all databases of Web of science. And the remaining 111,814 academic papers out of 116,193 academic papers have an average of 29.787 citations in Web of science core database and 32.755 citations in Web of science all databases. The average number of citations in Web of science core databases and the average number of citations in Web of science all databases of academic papers cited by the policy are 3.720 times and 3.742 times higher than that of academic papers that are not cited by the policy, respectively. As shown in Figure 7, Figure 8 below.
At the same time, we do Mann-Whitney U-test on the number of citations in WOS core database and WOS all databases for academic papers cited by the policy and academic papers not cited by the policy, and the results are shown in Fig. 9 and Fig. 10 below. Separate results, in the Mann-Whitney U test for the number of citations in the WOS core database, the U-Statistic is 378287334.5 with a p-value of 0.0. $p < 0.05$, indicating that there is a significant difference in the distribution of the number of citations to the two groups of academic papers cited by the policy and those not cited by the policy in the WOS core database, i.e. whether or not being cited by the policy has a significant effect on the number of citations of papers in the core database.
The box plot shows that the median number of citations of policy-cited papers is significantly higher than that of uncited papers, and the distribution is wider. While the number of citations for papers not cited by the policy is lower. And there are similar results in the Mann-Whitney U-test of the number of citations in all WOS databases. the U-Statistic is 378163246.0, and the P-value is 0.0. $p < 0.05$, indicating that there is a significant difference in the distribution of the number of citations to the two groups of policy-cited and non-policy-cited academic papers in all the databases of WOS. That is, whether or not they are cited by the policy has a significant effect on the number of citations of papers in the core databases. Observation of the box-and-line plot shows that the median number of citations in WOS all databases for policy-cited papers is significantly higher than that for uncited papers, and the distribution of citations is wider and contains more high citation values.The results of Yin,Gao's study suggest that policy documents about the COVID-19 pandemic substantially cite high-impact scientific results (Yin et al., 2021c). Although the data in our study are not as significant as in Yin, Gao's study, we still believe that our results also illustrate that policy documents in the supply chain field actually cite high-impact academic papers in the supply chain field.

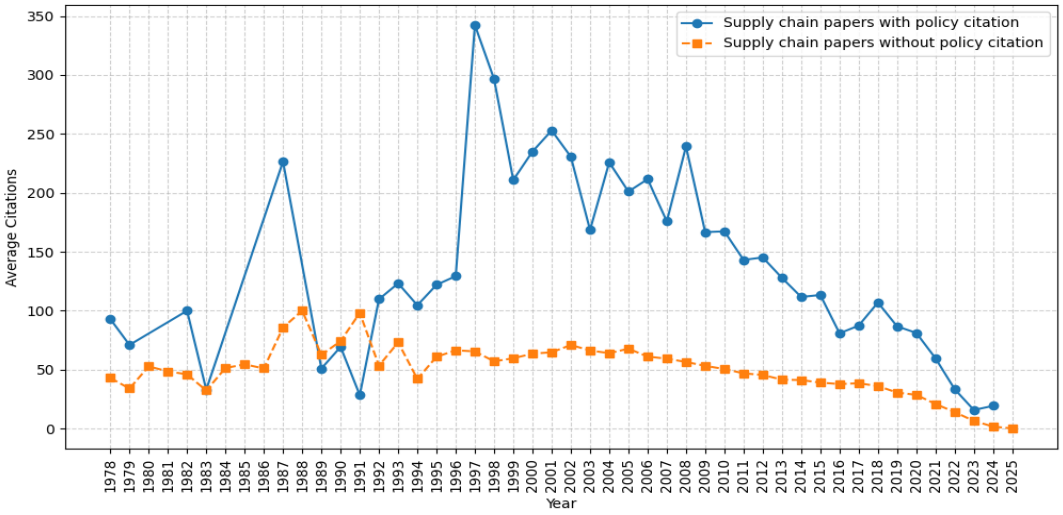


Figure 7. Difference in the number of core database citations between academic papers cited by the policy and those not cited by the policy.

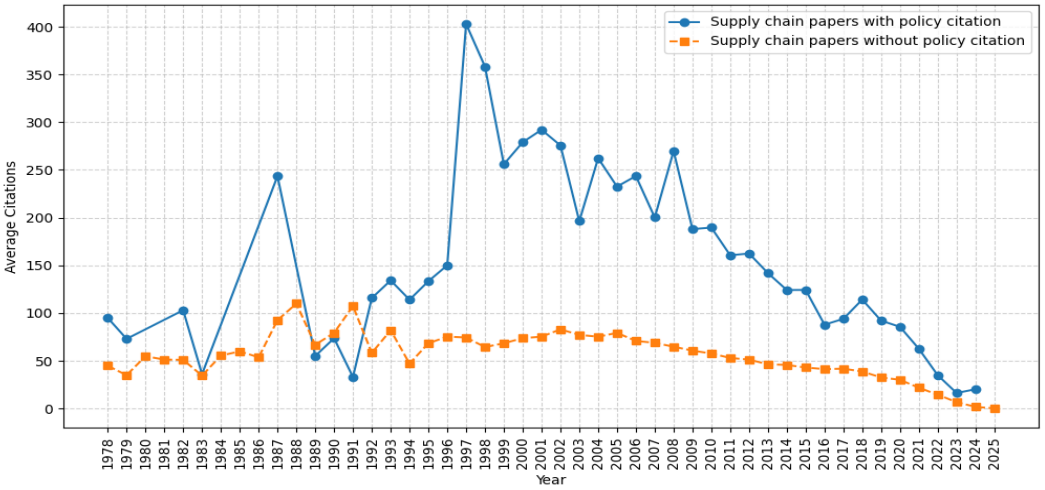


Figure 8. Difference in the number of citations in all databases between academic papers cited by the policy and those not cited by the policy.

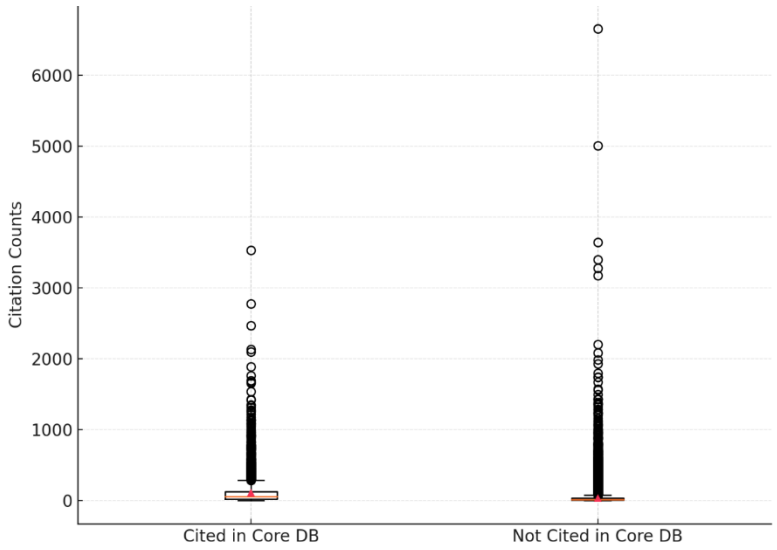


Figure 9. Comparison of the number of citations of policy-cited and non-policy-cited papers in the WOS core database.

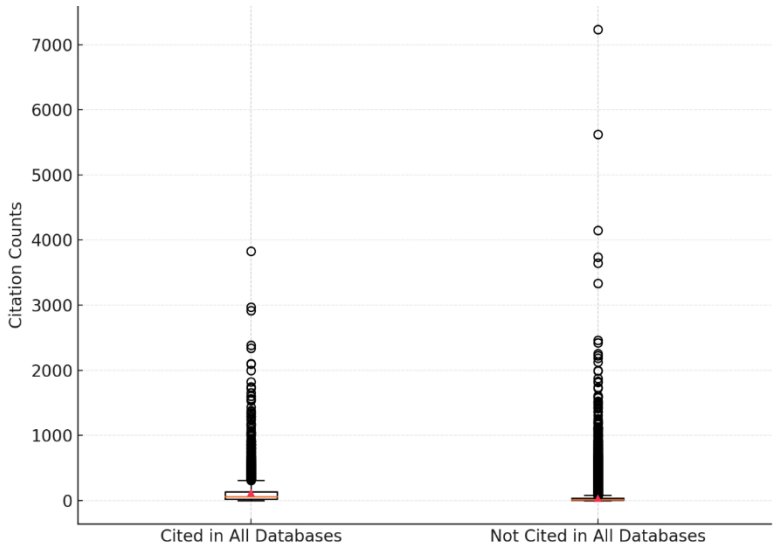


Figure 10. Comparison of the number of citations of policy-cited and non-policy-cited papers in the WOS all database.

Similarly, processing the acquired policy document data yields that the 8556 policy documents citing academic papers that are present in both policy collections have an average number of citations by policy after removing citations from their own policy source institutions of 8.667 citations, while the average number of citations by policy

that include citations from their own policy source institutions is 11.711 citations. The average number of policy citations for the remaining 229,293 policy documents out of 237,849 is 1.999 after removing citations from the same source, and the average number of policy citations for those that include citations from the same source is 3.158. (The number of policy citations we obtained from the Overtons database is divided into two categories, including citations from other policy documents from the policy document's own source institution and removing citations from other policy documents that are cited by the same source. source). The average number of policy citations after removing same-source citations and the average number of policy citations including same-source citations for policy documents that cite academic papers are 4.335 and 3.708 times higher than the average number of policy citations for policy documents that do not cite academic papers, respectively. Our findings also illustrate that supply chain policy papers that cite science also have higher citation levels in the supply chain policy domain. Therefore, we conclude that in the supply chain field, academic papers cited by policy papers are high level research papers in their own field, and policy papers that cite academic papers become high impact policies in their own field.

Conclusion

As mentioned earlier, the use of research results and recommendations of supply chain academic papers may be reflected in the citations of supply chain academic papers in policy documents. This study focuses on exploring the connection between scientific research and policy in the field of supply chain based on data from the Overton database and the Web of Science database, including policy documents and their citations to academic research papers, as well as the citations of policy documents and papers in their respective fields.

We can draw the conclusion that academic research improves the scientific nature of policies from the two-way interactive relationship between scientific research and policy making. Scientific research provides a rigorous theoretical basis and methodological tools for policy making. Academic research in the field of supply chain can provide a scientific basis for policy making by providing theoretical frameworks, data analysis and cutting-edge technological achievements. Policies that cite scientific research papers are more scientific in comparison. The high-level research papers cited in policy documents make these policies more authoritative in the field of supply chain, so they will be cited by more other policies and have a higher influence. By conducting quantitative analysis on academic research cited in policy documents, we can evaluate the actual impact of these studies on policy making and implementation, and then provide feedback for academic research in the field of supply chain and promote further optimization of research results. The new

results of academic research will once again promote the development of policies and maintain a good ecology of scientific research and policy making. In addition, some policy documents may refer to ideas, data or research findings in academic papers in their content, but these academic papers are not explicitly cited as sources in the text. This situation may be due to differences in the writing habits, length limitations or citation requirements of the policy documents.

The conclusions of this study point to the fact that improving the scientific quality and transparency of academic citations in policy documents, i.e., policymakers clearly citing the sources and rationale of academic research in policy documents, can improve the scientific quality and number of citations of policies, and increase the impact of policies. Publicly cited academic results in policy documents, when seen by academic researchers, can also promote understanding and support for the policy context within the academic community, again facilitating the synchronisation of scientific research and policy formulation.

At the same time, policy documents that cite academic research provide academic circles with cases where research results have been implemented, thus enhancing the practical application value of academic research. This shows that policies can also provide feedback to promote the deepening of academic research. In addition to policies assisting the implementation of academic research results, the focus on practical issues during the policy-making process will also drive the direction of academic research. For example, changes in prevention and control policies during the COVID-19 pandemic have promoted research on the stability of global supply chains, while regional economic development policies have promoted research on the localization and regionalization of supply chains. The citation of policy documents not only provides application scenarios for academic research, but also arouses researchers' attention to emerging issues and forms new theoretical and practical explorations. The academic community should encourage academic research to pay attention to policy needs, enhance sensitivity to policy needs, and pay attention to practical issues in policy making. In response to the current trend of globalization and regionalization of supply chains, relevant academic research should be carried out to provide timely support for policy adjustments.

In future research, a cooperation mechanism between academic research and policy making should be established. Policy-making departments and academic institutions should strengthen cooperation to achieve an effective combination of research and policy through joint research, policy consulting, etc. Future supply chain research needs to be more closely integrated with supply chain-related scientific research to enhance the wide applicability and policy influence of research results.

In summary, in the field of supply chain, a close two-way interactive relationship has been formed between high-level academic research and high-impact policy

documents. Academic papers cited by policy documents are high-level research papers in their own fields, and policy documents that cite academic papers have also become high-impact policies in their own fields. Academic research provides a scientific basis for policy making by providing theoretical foundations and technical support; policy documents enhance their authority by citing academic achievements, while promoting academic research to focus on practical problems. This virtuous circle not only enhances the scientificity and practicality of supply chain management, but also provides a guarantee for the effectiveness of policy making and implementation. In the future, by strengthening the cooperation mechanism between academia and policy, promoting the quantitative research of policy citations of scientific research, and exploring new directions for the integration of supply chain policy making and scientific research, we will make greater contributions to the sustainable development of the global economy and society.

Acknowledgments

This paper is one of the research outcomes of the National Social Science Fund of China project "Evaluation of Humanities and Social Sciences Academic Monographs Based on Scientometric Big Data" (Project No. 21BTQ104).

References

Ali Abd Al-Hameed, K. (2022). Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, 13(1), 3249-3255.

Bornmann, L., Haunschild, R., Boyack, K., Marx, W., & Minx, J. C. (2022). How relevant is climate change research for climate change policy? An empirical analysis based on Overton data. PloS one, 17(9), e0274693.

Brandts-Longtin, O., Lalu, M. M., Adie, E. A., Albert, M. A., Almoli, E., Almoli, F.,.. & Cobey, K. D. (2022). Assessing the impact of predatory journals on policy and guidance documents: a cross-sectional study protocol. BMJ open, 12(4), e059445.

Cabral, B., & Salles-Filho, S. (2024). Mapping science in artificial intelligence policy development: formulation, trends, and influences. Science and Public Policy, 51(6), 1104-1116.

Cheng, X., Tang, L., Zhou, M., & Wang, G. (2021). Coevolution of COVID-19 research and China's policies. Health research policy and systems, 19, 1-16.

Chowdhury, P., Paul, S. K., Kaisar, S., & Moktadir, M. A. (2021). COVID-19 pandemic related supply chain studies: A systematic review. Transportation Research Part E: Logistics and Transportation Review, 148, 102271.

Cristofoletti, E. C., Salles-Filho, S., Hollanda, S., Juk, Y., Pinto, K. E., Toledo, C. G.,.. & Campgnolli, E. (2023, April). The use of research in policy documents: exploring

methodological potentialities. In 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). International Conference on Science, Technology and Innovation Indicators.

Dorta-González, P., Rodríguez-Caro, A., & Dorta-González, M. I. (2024). Societal and scientific impact of policy research: A large-scale empirical study of some explanatory factors using Altmetric and Overton. Journal of Informetrics, 18(3), 101530.

Drongstrup, D., Malik, S., Aljohani, N. R., Alelyani, S., Safder, I., & Hassan, S. U. (2020). Can social media usage of scientific literature predict journal indices of AJG, SNIP and JCR? An altmetric study of economics. Scientometrics, 125, 1541-1558.

Fang, Z., Costas, R., Tian, W., Wang, X., & Wouters, P. (2020). An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics. Scientometrics, 124(3), 2519-2549.

Haunschild, R., & Bornmann, L. (2017). How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. Scientometrics, 110, 1209-1216.

Haunschild, R., Williams, K., & Bornmann, L. (2023, April). How relevant is public policy and administration research for the policy sector? An empirical analysis based on Overton data. In 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). International Conference on Science, Technology and Innovation Indicators.

Hodges, R., Caperchione, E., Van Helden, J., Reichard, C., & Sorrentino, D. (2022). The role of scientific expertise in COVID-19 policy-making: evidence from four European countries. Public Organization Review, 22(2), 249-267.

Hu, L., Huang, W. B., & Bu, Y. (2024). Interdisciplinary research attracts greater attention from policy documents: Evidence from COVID-19. *Humanities and Social Sciences Communications*, *11*(1), 1-10.

Huang, Z., Zong, Q., & Ji, X. (2022). The associations between scientific collaborations of LIS research and its policy impact. Scientometrics, 127(11), 6453-6470.

Hui, A., Rains, L. S., Todd, A., Boaz, A., & Johnson, S. (2020). The accuracy and accessibility of cited evidence: a study examining mental health policy documents. Social psychiatry and psychiatric epidemiology, 55, 111-121.

Ji, C. Y., Tan, Z. K., Chen, B. J., Zhou, D. C., & Qian, W. Y. (2024). The impact of environmental policies on renewable energy investment decisions in the power supply chain. Energy Policy, 186, 113987.

Jonker, H., & Vanlee, F. (2024). Linking science with media and policy: The case of academics in Flanders, Belgium. Quantitative Science Studies, 5(3), 556-572.

Llewellyn, N. M., Weber, A. A., Pelfrey, C. M., DiazGranados, D., & Nehl, E. J. (2023). Translating scientific discovery into health policy impact: innovative bibliometrics bridge translational research publications to policy literature. Academic Medicine, 98(8), 896-903.

Ma, J., & Cheng, Y. (2024). Why do some academic articles receive more citations from policy communities? Public Administration Review.

Marx, W., Haunschild, R., Thor, A., & Bornmann, L. (2017). Which early works are cited most frequently in climate change research literature? A bibliometric approach based on reference publication year spectroscopy. Scientometrics, 110, 335-353.

Nay, O., & Barré-Sinoussi, F. (2022). Bridging the gap between science and policy in global health governance. The Lancet Global Health, 10(3), e322-e323.

Newson, R., Rychetnik, L., King, L., Milat, A., & Bauman, A. (2018). Does citation matter? Research citation in policy documents as an indicator of research impact–an Australian obesity policy case-study. Health Research Policy and Systems, 16, 1-12.

Obuku, E. A., Sewankambo, N. K., Mafigiri, D. K., Sengooba, F., Karamagi, C., & Lavis, J. N. (2018). Use of post-graduate students' research in evidence informed health policies: a case study of Makerere University College of Health Sciences, Uganda. Health research policy and systems, 16, 1-13.

Overton (2020). Overton Help Center: Advice and answers from the Overton Team, Retrieved December 7, 2024 from: http://help.overton.io/en/ .

Pinheiro, H., Vignola-Gagné, E., & Campbell, D. (2021). A large-scale validation of the relationship between cross-disciplinary research and its uptake in policy-related documents, using the novel Overton altmetrics database. Quantitative Science Studies, 2(2), 616-642.

Rahimi, F., & Danesh, F. (2024). Scientometric Analysis of Political Documents of Overton: Open Government Data Case Study. Caspian Journal of Scientometrics, 11(2), 1-13.

Ren, C., & Yang, M. (2023). Study on the Characteristics of Cross-Domain Knowledge Diffusion from Science to Policy: Evidence from Overton Data. Proceedings of the Association for Information Science and Technology, 60(1), 368-378.

Schnake-Mahl, A. S., Jahn, J. L., Purtle, J., & Bilal, U. (2022). Considering multiple governance levels in epidemiologic analysis of public policies. Social Science & Medicine, 314, 115444.

Szomszor, M., & Adie, E. (2022). Overton: A bibliometric database of policy document citations. Quantitative science studies, 3(3), 624-650.

Van Elsland, S. L., O'Hare, R. M., McCabe, R., Laydon, D. J., Ferguson, N. M., Cori, A., & Christen, P. (2024). Policy impact of the Imperial College COVID-19 Response Team: global perspective and United Kingdom case study. Health Research Policy and Systems, 22(1), 153.

Vilkins, S., & Grant, W. J. (2017). Types of evidence cited in Australian Government publications. Scientometrics, 113(3), 1681-1695.

Watson, R. T. (2005). Turning science into policy: challenges and experiences from the science–policy interface. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1454), 471-477.

Xu, C., & Zong, Q. (2023). The effects of international research collaboration on the policy impact of research: A causal inference drawing on the journal Lancet. Journal of Information Science, 01655515231174381.

Yang, C., Huang, C., & Su, J. (2020). A bibliometrics-based research framework for exploring policy evolution: A case study of China's information technology policies. Technological Forecasting and Social Change, 157, 120116.

Yin, Y., Gao, J., Jones, B. F., & Wang, D. (2021). Coevolution of policy and science during the pandemic. Science, 371(6525), 128-130.

Yoshida, Y., Sitas, N., Mannetti, L., O'Farrell, P., Arroyo-Robles, G., Berbés-Blázquez, M.,.. & Harmáčková, Z. V. (2024). Beyond academia: a case for reviews of gray literature for science-policy processes and applied research. Environmental Science & Policy, 162, 103882.

Yu, H., Cao, X., Xiao, T., & Yang, Z. (2020). How accurate are policy document mentions? A first look at the role of altmetrics database. Scientometrics, 125, 1517-1540.

https://doi.org/10.51408/issi2025_038

The Trends of Open Access Academic Books and Discipline Dynamics: A Cross-database Comparison Based on OpenAlex and Web of Science

Li Jiangbo[1], Niu Shihang[2], Ouyang Wenhao[3], Li Jian[4], Zhang Mingyue[5]

[1]*jiangbosky@sina.com* ,
Business College, Qingdao University, Qingdao 266100
Research Institute for Science and Technology System and Institutional Innovation of Qingdao, Qingdao 266100 (China)

[2] *15648652050@163.com,* [3] *2111903049@qq.com,*[4] *lj13181433265@163.com,*
[5] *zhangmingyue1@qdu.edu.cn*
Business College, Qingdao University, Qingdao 266100 (China)

Abstract

This study conducts a quantitative analysis of the inclusion of open access (OA) academic books in the OpenAlex and Web of Science (WoS) databases from 2004 to 2023. It explores the differences and trends between the two databases in terms of the number of OA books, annual variations, and disciplinary distribution. The study finds that OpenAlex shows a significant advantage over WoS in the scale of OA book inclusion and growth rate, with a particularly strong growth potential in recent years. Through an analysis of various academic disciplines, this study reveals the differences in disciplinary preferences and resource distribution between the two databases, further discussing the advantages and challenges of OpenAlex in promoting the inclusion of OA books. The results indicate that OpenAlex demonstrates significant potential in advancing the inclusion of OA books and global knowledge sharing. OpenAlex has gradually become an important academic resource platform, providing more convenient access to the academic community. Additionally, the study innovatively uses large language models to resolve inconsistencies in cross-database disciplinary classification, enhancing the efficiency and accuracy of data matching. In conclusion, OpenAlex's rapid growth and its advantages across multiple disciplines provide crucial support for the global inclusion and sharing of OA books, while also offering important empirical data for the optimization of academic publishing policies and resource distribution.

Introduction

Academic books are primarily focused on scholarly research, using systematic and in-depth methods to explore significant academic achievements in specific fields or topics. Compared to academic papers, academic books provide deeper theoretical analysis and a more systematic framework. They offer an overall understanding and

thorough exploration of a research area, allowing for a more comprehensive and in-depth discussion of core issues within a discipline. As such, they often exert lasting influence in the academic community (Engels, 2018). Despite the increasing speed and impact of academic journals in recent years, academic books remain central in many disciplines, particularly in the humanities and social sciences. As the core medium for knowledge accumulation and academic communication, academic books carry substantial scholarly contributions and offer broader, more detailed perspectives. They serve as the foundation for building personal academic reputation and status (Zuccala et al., 2018; Kousha et al., 2018).

Open access (OA) is central to open science (OS), serving as a key framework for enhancing research transparency, reproducibility, and collaboration through practices like open data and open communication (Harnad, 2012). OA plays a key role by removing paywalls and increasing research accessibility, but its publishing model faces challenges such as high costs and selective accessibility, which may hinder the broader goals of OS (Pulverer, 2018). Despite these limitations, OA has significantly reshaped academic publishing by broadening knowledge dissemination and creating new opportunities (Zhang, 2024). Research also highlights its impact on citation diversity, as OA publications receive a wider range of citations than closed-access works (Huang et al., 2024) and are cited more frequently, particularly in recent publications (Yang et al., 2024). Additionally, technological advancements further optimize OA's role in research dissemination—artificial intelligence and machine learning enhance data processing and scientific discovery (Barbier et al., 2022), while OA articles tend to perform better in citation and alternative metrics, such as social media mentions, with green OA showing similar advantages (Clayson et al., 2021). In sum, OA is central to OS, driving more accessible and impactful research, but realizing its full potential requires addressing cost, accessibility, and quality concerns.

In the context of open science, OpenAlex, as an open academic platform, has become an important tool for academic research and data analysis due to its support for open access data, free API, and various query methods (Velez-Estevez et al., 2023; Delgado-Quirós et al., 2024; Hazarika et al., 2024; Harder et al., 2024). Through comparative research across multiple databases, Akbaritabar et al. (2023) found that OpenAlex excels in the scope and update frequency of journal inclusion, becoming an important resource for researching the latest academic achievements. Scheidsteger et al. (2023), by comparing the metadata of OpenAlex and the Microsoft Academic Graph (MAG), pointed out that both databases show a high degree of consistency in data, with OpenAlex making improvements in specifying document types, thus enhancing its value for bibliometric analysis. Aria et al. (2024) introduced OpenAlex's R package (openalexR), providing researchers with more

efficient and convenient analytical tools. However, OpenAlex also faces some challenges. Zhang et al. (2024) noted that over 60% of journal articles lack institutional information, particularly in early literature in the social sciences and humanities. The study recommends improving data quality and reducing research bias by supplementing missing data and strengthening collaboration among platforms, publishers, and users. Overall, OpenAlex has become an important resource in academic research due to its openness and diverse data support.

At present, research on academic databases mainly focuses on academic papers, particularly in areas such as literature indexing, classification, and bibliometric analysis. In contrast, research on academic books remains scarce, especially concerning the issue of OA books. Existing studies primarily concentrate on various characteristics of academic papers, such as the accuracy of literature indexing between databases (Jiao et al., 2023), the coverage of retracted literature (Ortega, 2024), differences in academic paper classification across disciplines (Singh, 2020), and the proportion of open access literature (Basson, 2022). Therefore, despite the importance of academic books as a form of scholarly communication, their distribution and influence in databases have not received sufficient attention. Web of Science, as one of the world's leading academic literature databases, is widely used in academic research and covers a vast amount of journal and conference literature. However, research on its inclusion and analysis of academic books, particularly open access books, remains relatively limited. This study aims to fill this gap by comparing the inclusion of OA academic books in WoS and OpenAlex across different disciplines, analyzing the differences and trends in OA book inclusion and disciplinary distribution between the two databases. The goal is to provide the academic community with a new perspective on the dissemination and development of OA academic books across disciplines and offer data support and theoretical foundations for future policy development and the optimization of academic resources.

Data Acquisition and Data Processing

Data Sources and Initial Processing

This study selected the metadata of all open access academic books published between 2004 and 2023 from the OpenAlex and WoS databases as the foundational data (retrieval date: December 15, 2024). In the data processing phase, manual filtering was first applied to exclude records that lacked disciplinary classification fields or publication dates. After filtering, the valid data records in the OpenAlex database amounted to 255,810, while the WoS database contained 8,713 valid records. Given the large volume of data in OpenAlex, this study utilized OpenAlex's

cursor mechanism via its API to continuously request the OA book metadata that met the criteria, ensuring the completeness of the records. For the WoS database, due to its relatively smaller dataset, the relevant metadata was directly exported through the WoS official website.

This study extracts information on OA books based on metadata from the OpenAlex and WoS databases. OpenAlex identifies OA status using the "open_access_is_oa" field, a Boolean variable ("true" or "false") indicating whether a book is OA. In contrast, WoS assigns OA status through the "Open Access Designations" field. While both databases provide classification methods for OA books, the relatively limited number of OA books indexed in WoS may result in an insufficient sample size when further distinguishing OA types, such as gold and green OA. This limitation could affect the robustness of statistical analyses. Therefore, this study adopts a binary classification (OA vs. non-OA), focusing on the overall inclusion of OA books without differentiating specific OA models. Additionally, this approach ensures comparability across databases, thereby enhancing the reliability of the findings.

Consideration and Evaluation of DOAB as a Benchmark

At the initial stage of this study, we explored the possibility of using the Directory of Open Access Books (DOAB) as a benchmark database to evaluate the coverage of OA books in WoS and OpenAlex. DOAB, established by the OAPEN Foundation in 2012, is a non-profit platform dedicated to indexing peer-reviewed academic books published under open access models (Maginiot et al., 2019). It has played an important role in the global open access publishing ecosystem and is often used in research related to OA policy and scholarly communication.

Our original intention was to treat DOAB as a comprehensive reference collection, enabling a comparative assessment of the OA book coverage between WoS and OpenAlex. However, after data retrieval and preprocessing, we found that DOAB's coverage in certain years was significantly lower than that of OpenAlex. For instance, in 2004, DOAB recorded approximately 700 OA books, whereas OpenAlex contained over 5,000 records for the same year. This considerable gap suggests that DOAB cannot serve as a stable benchmark for cross-database comparison. While WoS had an even lower number of OA books (around 50) in the same year, its indexing scope and selection criteria differ markedly from those of DOAB, further complicating the establishment of a unified standard.

Moreover, we encountered practical limitations related to metadata structure during the matching process. Metadata exported from DOAB—either through its API or web interface—generally lacks standardized identifiers such as ISBNs, which makes precise, record-level matching unfeasible.

Metadata exported from DOAB—whether via its API or web interface—lacks ISBN information, making precise one-to-one book-level matching infeasible. In large-scale data processing scenarios, fuzzy matching based on book titles alone introduces considerable uncertainty in accuracy and requires substantial computational resources and complex algorithmic support. We experimented with several text-matching techniques on a subset of records, but the results exhibited significant inconsistencies due to variations in naming conventions, language differences, and the handling of subtitles. These issues further undermined the stability of using DOAB as a reference dataset.

In summary, while DOAB remains a valuable initiative in promoting open access books and continues to be a key player in the OA ecosystem, its current limitations in data coverage, metadata completeness, and technical interoperability prevent it from serving as a reliable benchmark for evaluating WoS and OpenAlex in this study. Nevertheless, as DOAB continues to develop and enhance its data infrastructure, it holds promising potential for future OA-related bibliometric analyses.

Disciplinary Classification Normalization Using ChatGPT

When comparing the disciplinary fields of OA academic books between WoS and OpenAlex databases, the issue of disciplinary classification consistency emerged as a critical challenge. Different databases employ distinct disciplinary classification systems, which may lead to discrepancies during cross-database comparisons (Singh et al., 2021). Therefore, ensuring data consistency and comparability, particularly in standardizing disciplinary classifications, is crucial for this study. OpenAlex is an open academic platform based on the MAG, and its disciplinary classification system closely aligns with MAG (Priem et al., 2022). MAG, an academic graph created by Microsoft, utilizes a widely applied and systematic classification framework across multiple fields, with high academic data coverage (Sinha et al., 2015). In contrast to WoS's traditional classification system, OpenAlex and MAG offer a more simplified and systematic classification approach, categorizing academic research into four major domains and 26 specific fields. This classification method not only facilitates cross-disciplinary categorization but also better accommodates the demands of big data and diverse academic resources, particularly in the inclusion and classification of OA academic resources. Compared to traditional databases, OpenAlex provides a more open, flexible, and comprehensive classification framework, effectively supporting the systematic classification and analysis of OA resources. Thus, this study adopts OpenAlex's classification system as the standard to ensure the accuracy and consistency of cross-database comparisons.

Currently, several solutions to the issue of disciplinary classification standardization have been proposed. In Gao et al. (2024), educational discipline classifications were

manually mapped to the MAG research discipline classification system, successfully linking education disciplines to research disciplines, which provided strong support for analyzing the relationship between AI education and research. Sīle et al. (2021) used cross-mapping tables to map categories from different classification systems to the OECD R&D field classification system, eliminating discrepancies between classification systems and ensuring consistency and comparability of cross-system data, thus improving the reliability of the results. Osmani et al. (2023) proposed an improved method combining recursive grouping, clustering, and classification techniques to enhance disciplinary classification consistency, especially when facing complex classification systems, providing more accurate and stable classification results. Furthermore, the ECHO project created by Wittenburg et al. (2004) effectively solved the issue of inconsistent cross-disciplinary metadata by establishing a unified ontology structure and mapping relationships, and ensuring the interoperability of metadata via XML, providing important support for cross-database data comparison.

To resolve the differences between the disciplinary classification systems of WoS and OpenAlex, this study introduces Large Language Model (LLM) technology. As one of the most important advancements in artificial intelligence, LLMs have demonstrated powerful capabilities in solving various tasks in natural language processing, particularly with the emergence of ChatGPT-4, which has had a significant impact on AI development (Zhao et al., 2023). These models' exceptional performance in natural language processing makes them a valuable tool for solving complex tasks. ChatGPT exhibits strong semantic understanding, enabling it to handle complex contextual information and multidimensional classification problems effectively. Research has shown that ChatGPT's scientific feedback generation closely aligns with human responses, demonstrating robust critical thinking abilities (Liang et al., 2024). De Winter (2023) experimentally demonstrated that ChatGPT not only outperformed human annotators in annotation tasks but also exhibited superior language proficiency and reasoning, surpassing traditional cognitive models. This opens up new possibilities for disciplinary classification mapping.

Specifically, this study employs a custom Python script to systematically submit the *WoS Categories* metadata of 8,713 OA academic books from the WoS database to OpenAI's API. Using the ChatGPT-4o model, these categories are mapped to the *Field* metadata in the OpenAlex database. We interact with the model using the following prompt: "*You are an academic book disciplinary field mapper. The following academic discipline from the Web of Science database needs to be mapped to one of the 26 major fields in the OpenAlex database. Please classify it into the most appropriate OpenAlex field from the following list: {', '.join(openalex_fields)}.*

Return only the field name without any explanation. WoS Category: {wos_category}". The *wos_category* variable contains disciplinary classification data extracted from an Excel spreadsheet, representing the WoS subject categories assigned to individual books. The *openalex_fields* variable stores OpenAlex's standardized taxonomy of 26 disciplinary fields as a list-type data structure. This process relies exclusively on the discipline names provided by WoS, eliminating potential interference from other information and thereby ensuring the reliability and consistency of the classification. To enhance the stability of the results, the temperature of the ChatGPT-4o model is set to 0, ensuring minimal output variability and further improving the accuracy of subject mapping (DE, 2024). Table 1 presents the mapping relationships for selected representative disciplinary fields.

Table 1. WoS-OpenAlex Disciplinary Field Mapping (Selected).

No	*OpenAlex Disciplinary Field*	*WoS Disciplinary Field*
1	Agricultural and Biological Sciences	Entomology
		Food Science & Technology
		Ecology
2	Arts and Humanities	History & Philosophy of Science
		Literary Theory & Criticism
		Humanities
		Linguistics
3	Earth and Planetary Sciences	Meteorology & Atmospheric Sciences
		Geology
4	Engineering	Transportation
		Mechanics
5	Environmental Science	Water Resources
		Environmental Studies
6	Health Professions	Health Care Sciences & Services
		Psychiatry
		Medical Informatics

		Surgery
7	Immunology and Microbiology	Parasitology
		Microbiology
8	Physics and Astronomy	Thermodynamics
		Physics
9	Social Sciences	Sociology
		Information Science & Library Science
		Education
		Political Science

Method

This study employs a quantitative analysis method, combining statistical techniques and data visualization to deeply analyze the OA books data in the WoS and OpenAlex databases. First, we analyzed the annual changes in the proportion of OA books in both databases. Specifically, we extracted the number of OA books for each year from 2004 to 2023 for each database and calculated their proportion within the total number of acdemic books. The calculation formula is as follows:

$$P_t = \frac{OA_t}{T_t} \times 100\%$$

Where P_t represents the proportion of OA books in year *t*, OA_t is the number of OA books in year *t* in the database, and T_t is the total number of books in year *t* in the database. The variation in the proportion of OA books each year is presented in line graphs, allowing for an intuitive analysis of the OA book inclusion trends in these two databases. This will help us understand the development dynamics of OA books in both databases and assess which database is growing faster, thus projecting its future potential in OA book inclusion.

Secondly, we calculated the proportion of OA books in each disciplinary field for both the WoS and OpenAlex databases. This analysis examines the preference of each database in the inclusion of OA books in various disciplines. The formula used is as follows:

$$F_i = \frac{OA_i}{OA_{\text{Total}}} \times 100\%$$

Where F_i represents the proportion of OA books in the *i*-th disciplinary field relative to the total number of OA books in that database, OA_i is the number of OA books in

the i-th field, and OA_{Total} is the total number of OA books in the database. A higher proportion indicates a stronger inclination of the database to include OA books in that field, while a lower proportion indicates more limited inclusion of books in that field. Finally, we calculated the relative proportion of each disciplinary field, as follows:

$$D_i = \frac{X_i - Y_i}{X_i + Y_i}$$

Where D_i represents the relative proportion of OA books in the i-th discipline across the two databases, X_i is the proportion of OA books in the i-th field in the OpenAlex database, and Y_i represents the proportion of OA books in the same field in the WoS database. To visualize the results more intuitively, a heatmap was used. The color intensity in the heatmap reflects the relative proportion differences, allowing us to clearly identify the dominant fields in each database. By calculating the proportion of OA books in each disciplinary field and the differences in the relative proportions between fields, this study reveals the disparities in the inclusion of OA books across different disciplines between WoS and OpenAlex. This analysis not only provides data support for database optimization and resource allocation but also helps researchers in various fields select the most appropriate database for academic retrieval to improve research efficiency.

Results

Over the past two decades, the total proportion of OA books in the WoS database has been 6.01%, while the proportion in the OpenAlex database is 9.46%. Although this difference might not seem substantial in terms of overall proportion, considering the actual number of OA academic books in each database (OpenAlex includes 255,810 books, and WoS includes 8,713 books), it is clear that OpenAlex has significantly outpaced WoS in the volume of OA books. This indicates that OpenAlex has a broader scale and coverage in OA academic book inclusion. In other words, while the overall proportion difference is not highly significant, OpenAlex exhibits a stronger growth potential in the inclusion of OA books.

Next, this study further analyzes the annual changes in the proportion of OA books in both databases from 2004 to 2023. Figure 1 presents the line graph that visually illustrates the trend based on the proportion values for each year. It is evident that the proportion of OA books in the OpenAlex database has experienced significant growth over the past two decades. In 2004, the proportion of OA books in OpenAlex was 4.93%, and by 2023, it had dramatically increased to 25.10%, with a particularly sharp acceleration in growth after 2020. In contrast, while the proportion of OA books in the WoS database also shows an upward trend, the increase is relatively

slow. In 2004, the proportion of OA books in WoS was 6.00%, and by 2023, it had only increased to 11.32%. Notably, between 2004 and 2010, the proportion of OA books in WoS fluctuated significantly, and it consistently remained lower than OpenAlex. Although there has been some recovery in recent years, the growth rate remains considerably slower than OpenAlex.

Moreover, from the chart, it is evident that OpenAlex saw a significant increase in OA book proportion, from a low of 3.72% in 2011 to 25.10% in 2023, representing a growth of 576.34%. In contrast, WoS's proportion of OA books reached its lowest point of 3.74% in 2008 and only increased to 11.32% by 2023, a growth of 202.68%. This disparity in growth rates highlights the significant breakthrough OpenAlex has made in the inclusion of OA academic books.

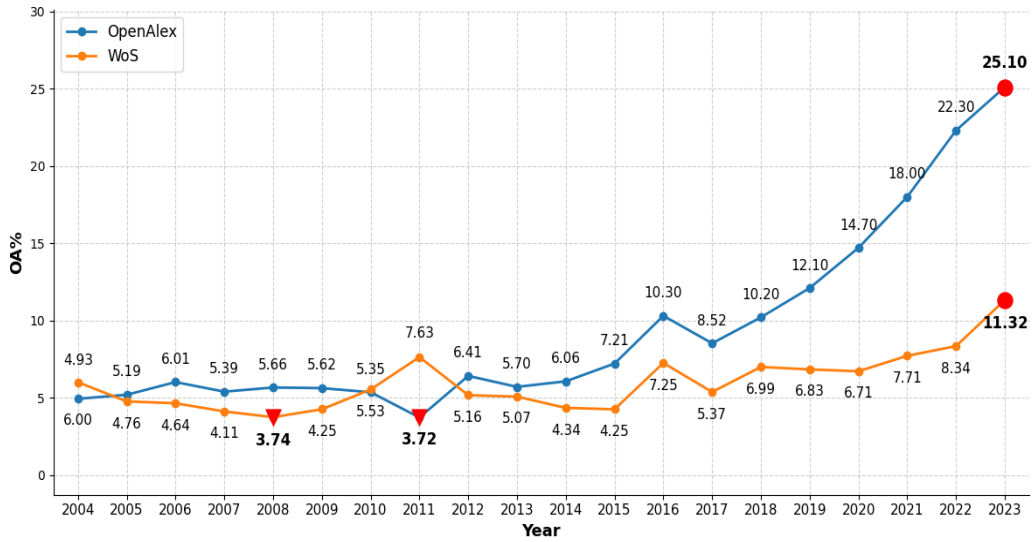


Figure 1. Trend of OA Book Proportions in WoS and OpenAlex (2004-2023).

Overall, OpenAlex's rate of growth in OA book inclusion is notably higher than WoS, demonstrating a stronger expansion potential. Particularly after 2018, OpenAlex's proportion of OA books quickly surpassed WoS and maintained a significant lead by 2023. This trend suggests that OpenAlex is likely to continue expanding its market share in the field of open access books in the coming years, while WoS may face the need to optimize its OA book inclusion strategy in response to the rapidly accelerating global open access trend.

Next, this study provides a detailed analysis of the 26 disciplinary fields, further subdivided based on the four major domains in the OpenAlex database (Life Sciences, Physical Sciences, Health Sciences, and Social Sciences). For each field, we calculated the proportion of OA books within that specific discipline to more accurately assess the inclusion of OA books in different fields. By independently

analyzing each field's sub-disciplines, this approach allows us to provide a more nuanced understanding of the inclusion of OA books within each respective domain.

In the Social Sciences domain, there are significant differences in the proportion of OA books between WoS and OpenAlex in several disciplines (Fig. 2A). OpenAlex shows higher proportions of OA books than WoS in the fields of *Social Sciences*, *Economics, Econometrics and Finance*, *Decision Sciences* and *Psychology*, with the largest difference in the *Social Sciences* field, where OpenAlex holds 53.61%, significantly higher than WoS's 35.20%. However, in the *Arts and Humanities* field, OpenAlex's proportion is 19.12%, much lower than WoS's 43.43%, indicating that WoS places greater emphasis on supporting and developing the humanities within the Social Sciences domain.

In the Physical Sciences domain, both WoS and OpenAlex exhibit significant advantages in different disciplines for the inclusion of OA books (Fig. 2B). In the fields of *Computer Science* and *Environmental Science*, OpenAlex's inclusion rate is significantly higher than WoS, likely reflecting its greater support for emerging disciplines. However, in the fields of *Mathematics*, *Chemistry*, *Materials Science* and *Physics and Astronomy*, WoS has a higher proportion of OA books, indicating that WoS has a stronger coverage in these traditional and foundational scientific fields, likely due to its longstanding authority and influence in scientific research. Additionally, the inclusion in other disciplines is relatively close between the two databases. For example, in the *Engineering* field, WoS's proportion is 27.76%, slightly lower than OpenAlex's 28.70%.

In the Life Sciences domain, WoS leads in the *Biochemistry, Genetics and Molecular Biology* field, with a proportion of 56.38%, significantly higher than OpenAlex's 31.25% (Fig. 2C). In contrast, OpenAlex shows stronger inclusion in the fields of *Agricultural and Biological Sciences* and *Neuroscience*, particularly in *Agricultural and Biological Sciences*, where OpenAlex's proportion is 54.73%, much higher than WoS's 35.90%. This difference indicates that OpenAlex has a stronger capability in the inclusion of OA books in emerging fields, and may continue to expand its inclusion in the Life Sciences domain. Furthermore, both databases show relatively weak inclusion in the *Veterinary* field, suggesting a need for further support and resource integration in this area.

Finally, in the Health Sciences domain, WoS has a significant lead in the *Health Professions* field, with a proportion of 84.73%, far exceeding OpenAlex's 38.53% (Fig. 2D). This indicates that WoS has stronger inclusion capabilities for OA books in health-related disciplines. However, in the *Medicine* field, OpenAlex has a proportion of 56.54%, while WoS is only at 9.36%. This disparity suggests that OpenAlex has a stronger capability in the inclusion of OA books in the medical sciences. In certain specialized subfields, such as *Pharmacology, Toxicology and*

Pharmaceutics and *Dentistry*, WoS shows higher OA book proportions of 3.20% and 2.22%, respectively, compared to OpenAlex's 0.99% and 0.93%. This indicates that WoS has stronger growth potential in specialized fields like pharmacology and dentistry.

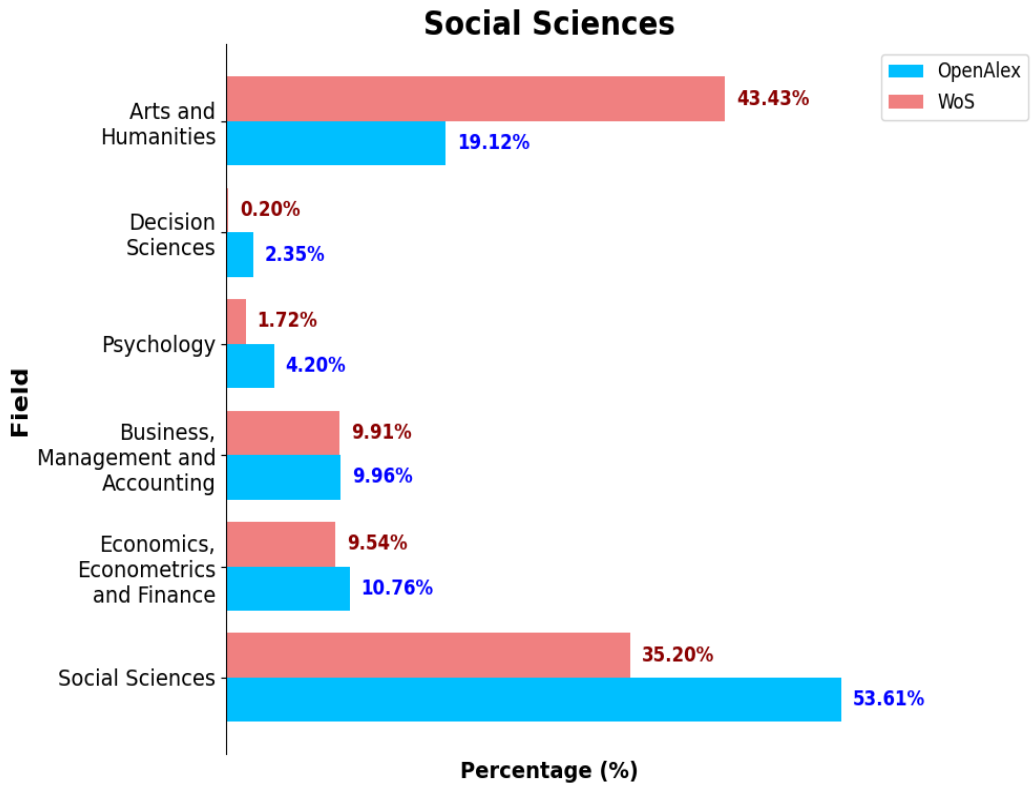


Figure 2A. Proportion of OA Books in Social Sciences by WoS and OpenAlex.

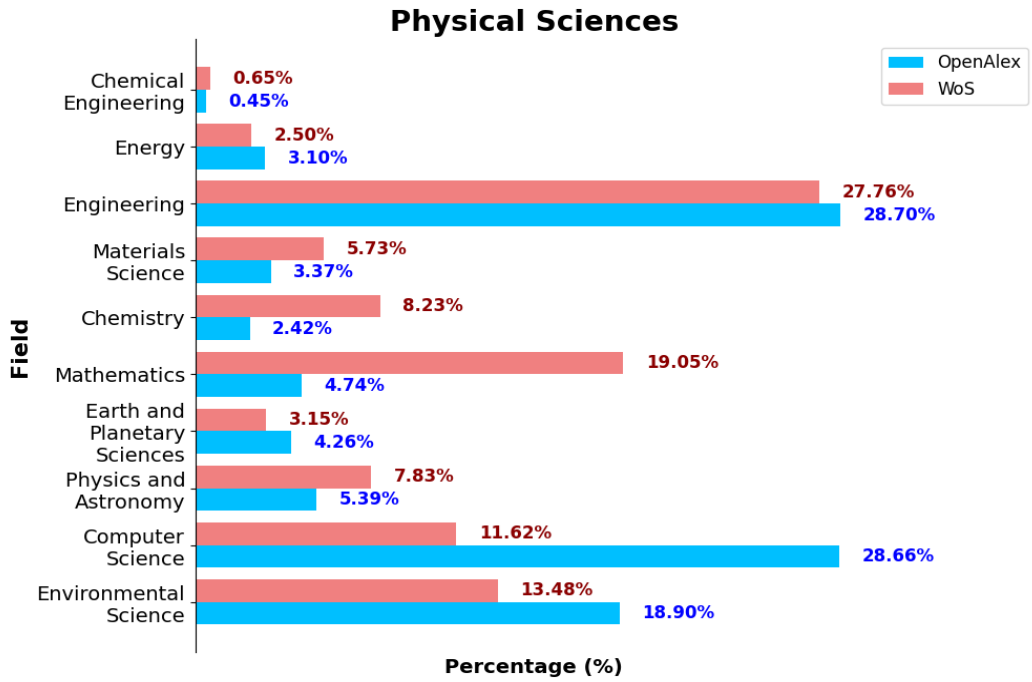


Figure 2B. Proportion of OA Books in Physical Sciences by WoS and OpenAlex.

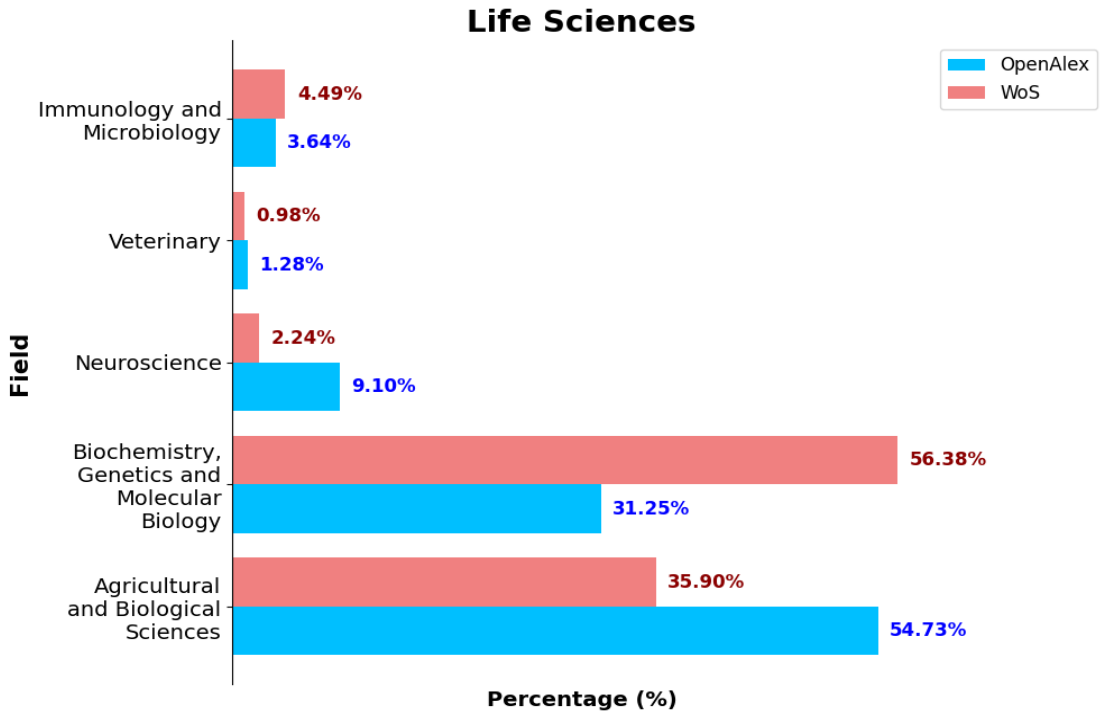


Figure 2C. Proportion of OA Books in Life Sciences by WoS and OpenAlex.

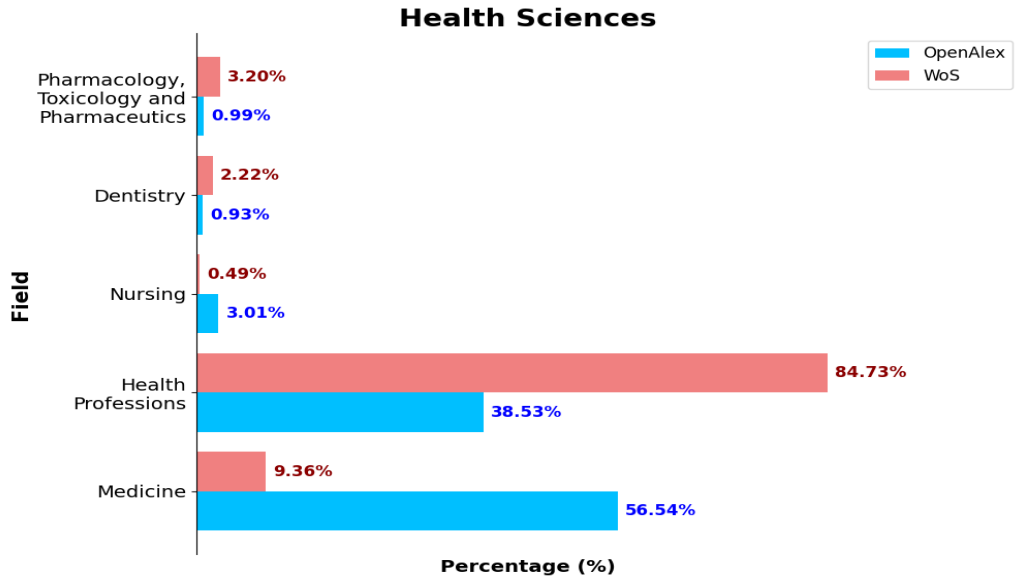


Figure 2D. Proportion of OA Books in Health Sciences by WoS and OpenAlex.

To visually represent the differences in the inclusion of OA books across disciplines in WoS and OpenAlex, this study calculated the relative proportions for each field. Figure 3 presents a heatmap of the relative proportions, providing a more intuitive visual representation of the databases' inclusion tendencies across different disciplines. In the heatmap, colors closer to red indicate a relative proportion near 1, suggesting that OpenAlex has a higher proportion of OA books in that field. Conversely, colors closer to blue represent a relative proportion near -1, indicating that WoS leads in that field, while shades of grey represent fields where the databases' inclusion is more similar. From the heatmap, it is apparent that a larger number of disciplines show a tendency towards red, indicating that OpenAlex includes more OA books in more fields.

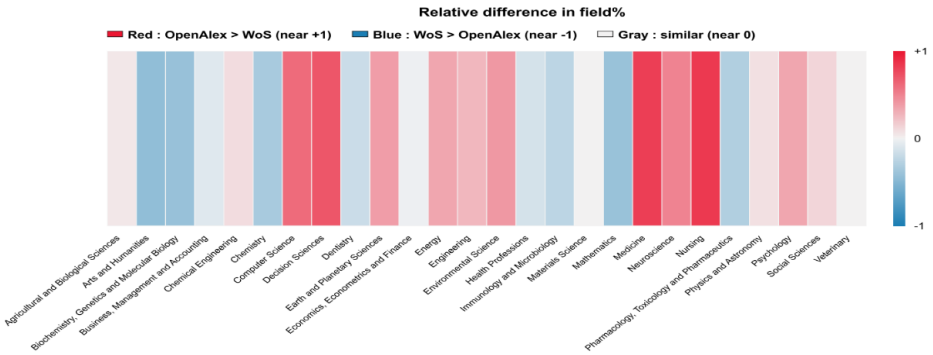


Figure 3. Heatmap of Relative Proportions of OA Book Inclusion in WoS and OpenAlex Across 26 Disciplinary Fields.

Discussion

This study employs quantitative analysis to examine the inclusion of OA academic books in the WoS and OpenAlex databases from 2004 to 2023, with a focus on comparing the number and distribution of OA books across different disciplines in both databases. The findings indicate that OpenAlex has a significantly higher number and growth rate of OA books than WoS, particularly in the past five years, during which the proportion of OA books in OpenAlex has increased substantially. This finding suggests that OpenAlex has the capacity to significantly contribute to the advancement of open-access book publishing.

The accuracy of discipline classification standardisation is a fundamental prerequisite for data analysis in this study. As WoS and OpenAlex employ different classification systems, their comparability and consistency are directly affected. To address these discrepancies, this study employs ChatGPT to match disciplinary classifications across the two databases, thereby ensuring a standardized classification system. Leveraging its advanced semantic understanding capabilities, GPT efficiently and accurately resolves inconsistencies between the classification systems of the two databases. Additionally, GPT serves as an auxiliary tool in this study, maintaining high accuracy while reducing human error and providing fast and reliable classification mapping. In comparison with conventional manual classification methodologies, GPT, as a large language model, not only enhances classification efficiency but also handles more complex and interdisciplinary

classification tasks, thereby significantly expanding the boundaries of academic data processing. The integration of GPT thus presents a novel approach to disciplinary classification and demonstrates the potential of artificial intelligence in academic research.

Moreover, this study conducts an in-depth analysis of OA books across a range of academic disciplines. The results reveal substantial differences in the proportion of OA books across disciplines in the two databases. A more pronounced advantage in the inclusion of OA books is exhibited by OpenAlex in fields such as *Computer Science* and *Environmental Science*, whereas a higher share of OA books in disciplines such as *Biochemistry, Genetics and Molecular Biology* and *Physics and Astronomy* is exhibited by WoS. This analysis provides valuable insights into the inclusion preferences of the two databases across various disciplines.

OpenAlex is a vital element of the open science ecosystem, offering a more open and sustainable model for academic resource sharing through its freely accessible API and extensive data coverage. In comparison to conventional subscription-based databases, OpenAlex boasts substantial advantages in terms of accessibility, openness, and interoperability. These qualities contribute to the reduction of inequalities in access to academic resources and the enhancement of global research collaboration. In light of the growing emphasis on open science policies, the development of OpenAlex is of paramount importance in promoting equity and transparency in academic publishing. Nevertheless, despite its strong potential in the inclusion of OA books, OpenAlex still faces certain limitations concerning metadata quality. For instance, the absence of institutional affiliation information for some books may affect the accuracy of author attribution and research impact analysis. In order to enhance its value in academic research, it is recommended that OpenAlex continue to improve metadata quality and enhance its integration with other open science tools.

This study provides a systematic analysis of the differences in OA book inclusion between WoS and OpenAlex, as well as empirical evidence for the formulation of open-access policies and academic resource management strategies. As open science continues to evolve, data interoperability and accessibility will become pivotal issues in global scholarly communication. The findings of this study highlight the role of OpenAlex in promoting the dissemination of OA books and offer valuable insights for the optimisation of open science infrastructure and the evolution of academic publishing models.

References

Akbaritabar, A., Theile, T., & Zagheni, E. (2023). *Global flows and rates of international migration of scholars* (No. WP-2023-018). Max Planck Institute for Demographic Research. https://doi.org/10.4054/MPIDR-WP-2023-018

Aria, M., Le, T., Cuccurullo, C., Belfiore, A., & Choe, J. (2024). openalexR: An R-tool for collecting bibliometric data from OpenAlex. *R Journal, 15*(4), 167-180.

Barbier, L. M., Green, J. L., & Draper, D. S. (2022). The need for open access and natural language processing. *Proceedings of the National Academy of Sciences, 119*(15), e2200752119.

Basson, I., Simard, M. A., Ouangré, Z. A., Sugimoto, C. R., & Larivière, V. (2022). The effect of data sources on the measurement of open access: A comparison of Dimensions and the Web of Science. *PLOS One, 17*(3), e0265545.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2021). The open access advantage for studies of human electrophysiology: Impact on citations and altmetrics. *International Journal of Psychophysiology, 164*, 103-111.

de Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics, 1-19*.

de Winter, J. C. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*, 1-16.

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies, 5*(1), 31-49.

Engels, T. C., Istenič Starčič, A., Kulczycki, E., Pölönen, J., & Sivertsen, G. (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? *Aslib Journal of Information Management, 70*(6), 592-607.

Gao, J., & Wang, D. (2024). Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nature Human Behaviour, 1-12*.

Harder, R. (2024). Using Scopus and OpenAlex APIs to retrieve bibliographic data for evidence synthesis: A procedure based on Bash and SQL. *MethodsX, 102601*.

Harnad, S. (2012). Open access: A green light for archiving. *Nature, 487*(7407), 302-303.

Hazarika, R., Roy, A., & Sudhier, K. G. (2024). Mapping the open access publications of Indian non-profit organizations over the last 20 years based on OpenAlex insights. *Global Knowledge, Memory and Communication*. https://doi.org/10.1108/GKMC-02-2024-0106

Huang, C. K., Neylon, C., Montgomery, L., Hosking, R., Diprose, J. P., Handcock, R. N., & Wilson, K. (2024). Open access research outputs receive more diverse citations. *Scientometrics, 129*(2), 825-845.

Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data, 10*(1), 737.

Kousha, K., & Thelwall, M. (2018). Can Microsoft Academic help to assess the citation impact of academic books? *Journal of Informetrics, 12*(3), 972-984.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., ... & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI, 1*(8), AIoa2400196.

Maginiot, F., Mounier, P., & Pellen, M. (2019). *DOAB Foundation: Toward a quality label for academic books in open access*. CNRS. https://www.cnrs.fr/en/press/doab-foundation-toward-quality-label-academic-books-open-access

Ortega, J. L., & Delgado-Quirós, L. (2024). The indexation of retracted literature in seven principal scholarly databases: A coverage comparison of Dimensions, OpenAlex, PubMed, Scilit, Scopus, The Lens, and Web of Science. *Scientometrics, 129*(7), 3769-3785.

Osmani, A., Hamidi, M., & Alizadeh, P. (2022). Clustering approach to solve hierarchical classification problem complexity. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7904-7912).

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Scheidsteger, T., & Haunschild, R. (2023). Which of the metadata with relevance for bibliometrics are the same and which are different when switching from Microsoft Academic Graph to OpenAlex? *Profesional de la Información, 32*(2). https://doi.org/10.3145/epi.2023.mar.09

Sīle, L., Guns, R., Vandermoere, F., Sivertsen, G., & Engels, T. C. (2021). Tracing the context in disciplinary classifications: A bibliometric pairwise comparison of five classifications of journals in the social sciences and humanities. *Quantitative Science Studies, 2*(1), 65-88.

Singh, P., Piryani, R., Singh, V. K., & Pinto, D. (2020). Revisiting subject classification in academic databases: A comparison of the classification accuracy of Web of Science, Scopus & Dimensions. *Journal of Intelligent & Fuzzy Systems, 39*(2), 2471-2476.

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus, and Dimensions: A comparative analysis. *Scientometrics, 126*, 5113-5142.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243-246).

Velez-Estevez, A., Perez, I. J., García-Sánchez, P., Moral-Muñoz, J. A., & Cobo, M. J. (2023). New trends in bibliometric APIs: A comparative analysis. *Information Processing & Management, 60*(4), 103385.

Wittenburg, P., Gulrajani, G., Broeder, D., & Uneson, M. (2004). Cross-disciplinary integration of metadata descriptions. In *LREC 2004*.

Yang, P., Shoaib, A., West, R., & Colavizza, G. (2024). Open access improves the dissemination of science: Insights from Wikipedia. *Scientometrics, 129*(11), 7083-7106.

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics, 1-23*.

Zhang, X. (2024). Is open access disrupting the journal business? A perspective from comparing full adopters, partial adopters, and non-adopters. *Journal of Scholarly Publishing, 55*(2), 145-162.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zuccala, A. A., Giménez-Toledo, E., & Peruginelli, G. (2018). Scholarly books and their evaluation context in the social sciences and humanities. *Aslib Journal of Information Management, 70*(6), 586-591.

https://doi.org/10.51408/issi2025_024

Trajectory of Research Method Usage in the Academic Careers of Scholars in the Library and Information Science

Jiayi Hao[1], Chengzhi Zhang[2]

[1]*haojiayi@njust.edu.cn*, [2]*zhangcz@njust.edu.cn*
Nanjing University of Science and Technology, Department of Information Management, Nanjing, 210094 (China)

Abstract

Research methodologies constitute an indispensable tool for scholars engaged in scientific inquiry. Investigating the trajectory of methodological usage throughout scholars' academic careers can illuminate distinctive patterns in their adoption of research methods, thereby offering valuable insights for novice researchers in selecting appropriate methodologies. This study employs a comprehensive dataset comprising full-text journal articles and bibliographic records from the Library and Information Science (LIS) domain. Utilizing an automated classification model based on full-text cognitive analysis, the research methods employed by LIS scholars are systematically identified. Subsequently, author name disambiguation is performed, and academic age is calculated for each scholar. The study focuses on a cohort of 435 senior scholars with an academic age exceeding 14 years and a consistent publication record at five-year intervals, encompassing a total of 6,116 articles. The findings reveal a trajectory in methodological selection characterized by an initial increase followed by a gradual decline over the course of scholars' careers. Furthermore, scholars exhibit a propensity for combining multiple research methods, including both conventional and unconventional pairings. Notably, the research methods most commonly used by researchers change with age and seniority.

Introduction

The scholarly endeavors of researchers serve as a driving force behind scientific progress. Investigating the characteristics of scholars themselves provides valuable insights into the mechanisms that shape modern science. Age, as a significant attribute of scholars, exerts a discernible influence on their academic performance. As scholars advance in age, their cognitive abilities and academic perspectives undergo expansion (Wang et al., 2017), thereby shaping their research interests, methodological choices, and the output of their scholarly contributions.

Given the unique and complex nature of academic research, prior studies have adopted the lens of academic age to more precisely delineate and comprehend the developmental trajectories and stage-specific characteristics of scholars within their respective fields. Academic age is typically calculated based on the timing of a scholar’s first publication (Costas et al., 2015). This metric has been extensively linked to various dimensions of scholarly activity, including research productivity (Abramo et al., 2016; Győrffy et al., 2020), academic influence (Sugimoto et al., 2016), and collaborative networks (Bu et al., 2018; Kumar & Ratnavelu, 2016; Wang et al., 2017). Understanding how scholars select and shift their research focus over time is of paramount importance, as it has implications for the training of scientists, the allocation of scientific funding, the organization and discovery of knowledge, and the recognition and reward of excellence (Jia et al., 2017). Academic age also

serves as a critical metric for distinguishing different stages of an academic career. Empirical studies reveal that as scholars progress in academic age, they accumulate greater resources and exhibit a heightened propensity to explore diverse research topics, accompanied by an increase in productivity (Abramo et al., 2016; Simoes & Crespo, 2020; Zeng et al., 2019). However, disparities exist between scholars of different ages. While senior scholars possess advantages in experience, funding, and collaboration, their knowledge base tends to stabilize in the later stages of their careers. This stabilization is often accompanied by the use of relatively outdated concepts (Liang et al., 2020; Milojević, 2012; Packalen & Bhattacharya, 2019), a diminished receptivity to novel ideas (Azoulay et al., 2019), and engagement in less prominent research areas (Cui et al., 2022). Consequently, scholars at different stages of their academic careers exhibit distinct cognitive behaviors and research patterns.

Research methods, as the cognitive frameworks guiding scientific inquiry, constitute an indispensable scientific element in the formation of any academic discipline. Serving as a cornerstone of scientific research, their significance and the urgency for innovation have become increasingly pronounced. Studies have revealed notable age-related differences in the research methods employed by scholars at various stages of their academic careers. Senior scholars exhibit a predilection for qualitative research, while their junior counterparts tend to favor quantitative methodologies (Lou et al., 2021). Consequently, there is a compelling need to explore the trajectory of methodological choices throughout scholars' academic careers. Previous research has predominantly examined the impact of academic age through the lenses of team collaboration, scholarly output, and related dimensions, or has focused on the classification, identification, and application of research methods. However, there is a notable gap in integrating academic age with the use of research methods to provide a comprehensive analysis of methodological evolution across the entirety of a scholar’s career. This oversight has led to the neglect of fundamental questions, such as what research methods scholars employ during their careers and the underlying logic and influencing factors driving these choices. Investigating the trajectory of methodological usage in scholars' academic careers can unveil distinctive patterns in their adoption of research methods, thereby offering valuable insights and guidance for early-career researchers in selecting appropriate methodologies for their scholarly pursuits.

This study employs journal literature as its primary data source to investigate the trajectory of research method usage among scholars in the Library and Information Science (LIS) domain, with a focus on individual scholars. The research aims to address the following questions:

RQ1: What differences exist in the research methods employed by LIS scholars at various stages of their academic age?

RQ2: What patterns characterize the trajectory of research method usage throughout the academic careers of LIS scholars?

Literature review

This paper aims to explore the trajectory of research method selection in the academic careers of scholars in a specific field. Given the relatively limited body of

research on scholars' academic trajectories, this study will focus on two key dimensions: academic age and the utilization of research methods.

Academic age of scholars in specific fields

Research on the academic age of scholars in specific fields can be divided into two main areas: the definition of academic age and the various dimensions of academic age research.

Regarding the calculation of academic age, existing studies predominantly rely on two metrics: the timing of a scholar's first publication and the year of doctoral graduation. However, the scale of these studies varies significantly. Research utilizing the first publication date to determine academic age encompasses a wide range of sample sizes. Smaller-scale studies span diverse fields, such as 137 scholars in information systems (Liao, 2017) and 472 top economists (Simoes & Crespo, 2020). Larger-scale studies include 21,562 scientists across five disciplines and ten core journals (Milojević, 2012), 94,000 scientists from 43 countries (Chan & Torgler, 2020), and even 222,925 authors (Robinson-Garcia et al., 2020) or 1.7 million author records from the Web of Science platform (Aref et al., 2019).

In contrast, studies using the year of doctoral graduation to calculate academic age typically involve smaller samples, often numbering in the hundreds (Badar et al., 2014; Chan & Torgler, 2020; Coomes et al., 2013) or thousands (Perianes-Rodriguez & Ruiz-Castillo, 2015; Sugimoto et al., 2016). For instance, van den Besselaar and Sandström (2016) examined 243 researchers applying for early-career grants in the Netherlands, while Perianes-Rodriguez and Ruiz-Castillo (2015) analyzed 2,530 economists working in 81 top global economics departments. Costas et al. (2015) utilized a real-world dataset of professors in Quebec to evaluate the feasibility of these two metrics and concluded that the first publication date is a more suitable indicator of a researcher's academic age. Similarly, Nane et al. (2017) identified the year of first publication as the best linear predictor of a scholar's age. Consequently, this study defines the starting point of a scholar's academic career as the timing of their first publication.

In research, academic age is often examined in conjunction with scholars' academic or professional trajectories and is explored from multiple perspectives, as illustrated in Table 1.

Table 1. Different research perspectives integrating scholars' academic careers.

Authors	*Perspective*	*Main findings*
Milojević (2012)	Reference citation behaviour	Similar citation behavior with senior and junior researchers citing references at comparable rates and consistent re-citation patterns
Aref et al. (2019)	Researcher mobility	Hypermobility analysis categorizing scholars at early mid and late career stages by academic age and identifying destination countries
Simoes and Crespo (2020)	Performance assessment	Publication productivity showing longer careers linked to higher output and prolific authorship
Robinson-Garcia et al. (2020)	Career trajectories	Career stage biases revealed through academic age and author contribution statements indicating variations in scientific trajectories
Ao et al. (2023)	Patterns of scientific creativity	Disruption index trends with both male and female scholars showing a "high peak" creativity pattern and a small subset of females exhibiting an "early peak"
Zhang et al. (2024)	Changes in research direction	Research direction shifts with women changing direction less frequently than men and experiencing less negative performance impact

It is evident that the use of academic age as an individual characteristic of scholars has matured significantly. This study integrates the metric of academic age to examine the trajectory of research methods employed by scholars at different stages of their academic careers.

Overview of research on the use of research methods in specific fields

Investigating and analyzing the use of research methods in academic papers can reveal and reflect the fundamental trends in the application and development of methodologies within a discipline. Table 2 summarizes studies on the use of research methods by scholars in the Library and Information Science field, highlighting diverse analytical perspectives. For instance, Järvelin and Vakkari (1990) categorized research methods in LIS core journal articles into nine research strategies and ten data collection methods. Chu (2015) classified LIS research methods into 16 categories based on data collection techniques. Hayman and Smith (2020) analyzed the use of mixed methods in articles, examining the extent of mixed methods research in LIS over the past decade (2008–2018) and the volume of such studies in health-related contexts. Additionally, some scholars have explored trends in the evolution of research methods. Lund and Wang (2021) employed visualization techniques to examine changes in the use of various research methods, finding that the diversity of methods used in articles has increased over time. Lou et al. (2021)

investigated how researchers in different age groups employ research methods over time. Järvelin and Vakkari (2021) expanded on their earlier work by summarizing the methodological evolution in LIS over the past 50 years, noting that LIS research has become increasingly methodologically diverse, with more varied approaches to analyzing research subjects. Zhang et al. (2023) conducted a longitudinal study on the frequency and diversity of research methods in LIS, revealing a shift from conceptual to empirical research strategies over 31 years.

In summary, the heightened attention scholars have paid to the use of research methods has contributed to the refinement of methodological paradigms within the field. However, few studies have integrated research methods with scholars' academic careers to explore their usage trajectories. Therefore, this study adopts a broader, dynamic perspective to investigate the evolution of research method selection throughout scholars' academic careers, uncovering the underlying mechanisms that drive these choices. This approach aims to provide valuable insights and recommendations for scholars regarding the application of research methods.

Table 2. Studies on the use of research methods.

Authors	*Perspective*	*Main findings*
Järvelin and Vakkari (1990)	Classification of research methods	Systematic categorization of research methods into 9 strategies and 10 data collection techniques
Chu (2015)	Classification of research methods	LIS research methods classified into 16 categories based on data collection
Hayman and Smith (2020)	Use of mixed research methods	Mixed methods in LIS showing small but significant growth over the past decade
Lund and Wang (2021)	Changing trends in the use of various research methods.	Increasing method diversity with data analysis and qualitative methods dominating recent publications
Lou et al. (2021)	Researchers in different age groups use research methods over time	Rise in quantitative methods driven by younger researchers and senior scholars
Järvelin and Vakkari (2021)	Research evolution in the field of LIS	Methodological fragmentation in LIS over 50 years reflecting diversified analytical approaches.
Zhang et al. (2023)	Frequency and diversity of application of research methods	Shift in LIS research strategies from conceptual to empirical over 31 years

Data and methodology

This section outlines the research framework and key steps for investigating the trajectory of research method selection among scholars in a specific field throughout

their academic careers. The framework encompasses data sources, classification of research methods, and the acquisition of academic career data for scholars in the specified field.

Framework

This study aims to explore the trajectory of research method selection in the academic careers of scholars in the LIS field. Firstly, full-text journal articles and related bibliographic records from the specified field serve as the primary data sources. Machine learning techniques are employed to identify research methods within these texts, enabling the construction of a comprehensive research method corpus for the field. Secondly, name disambiguation is carried out on the authors in the collection of academic papers, and information such as the academic age of scholars is calculated. Then, research on the trajectory of research method selection is conducted according to the relevant data of the selected senior scholars.
The research framework is shown in Figure 1.

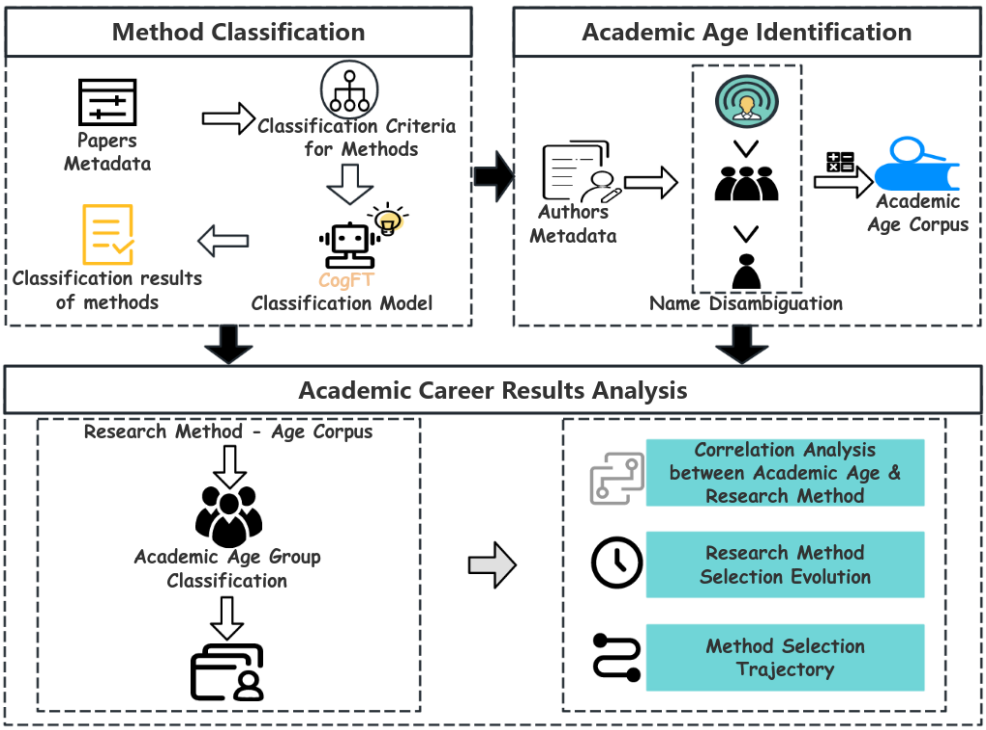


Figure 1. Framework of this study.

Data sources

The focus of this study is scholars in the Library and Information Science field, and the data sources are academic journals within this domain. In prior research, Järvelin and Vakkari (1993) conducted extensive studies on research methods and identified 31 representative academic journals in LIS based on the research topics covered in their articles. Building on this foundation, this study integrates the list of

representative journals identified by Järvelin and colleagues with the 2023 Journal Citation Reports (JCR) LIS category, which includes core journals across quartiles Q1 to Q4. This process resulted in the selection of 14 high-quality, representative LIS journals. Consequently, the full-text data collection for this study encompasses scholarly articles published in these 14 high-quality LIS journals. The data types collected include both metadata and full-text data, covering the period from 1990 to 2023. Full-text data were obtained from the official websites of each journal and converted into Word document format using conversion tools. These documents were then processed and parsed using Python to generate standardized full-text data. For cases where metadata were incomplete, bibliographic data for all articles published in the 14 journals over the 34-year period were downloaded from the Web of Science (WoS) [https://www.webofscience.com], and missing metadata were supplemented using DOI matching. In total, this study compiled full-text and metadata for 26,677 academic articles published in LIS journals between 1990 and 2023. The number of articles per journal is detailed in Table 3.

Table 3. Number of academic articles in high quality representative journals in the field of LIS.

No.	*Journal name*	*Abbreviation*	*Number of Articles*
1	*Aslib Journal of Information Management*	AJIM	1356
2	*College & Research Libraries*	CRL	1330
3	*Information Processing & Management*	IPM	3063
4	*Information Technology and Libraries*	ITL	546
5	*International Journal of Information Management*	IJIM	1891
6	*Journal of Documentation*	JOD	1450
7	*Journal of Information Science*	JIS	1510
8	*Journal of Librarianship and Information Science*	JLIS	887
9	*Journal of the Association for Information Science and Technology*	JASIST	3928
10	*Library & Information Science Research*	LISR	783
11	*Library Quarterly*	LQ	502
12	*Online Information Review*	OIR	1684
13	*Scientometrics*	SCIM	5926
14	*Electronic Library*	TEL	1821

Among the 14 journals, the three journals with the highest number of data entries are *Scientometrics*, *Journal of the Association for Information Science and Technology*, and *Information Processing & Management*. These journals collectively account for 12,917 articles, representing nearly 50% of the total dataset.

Classification of research methods for academic papers in the LIS field

Based on the constructed full text corpus of academic papers in the field of LIS, this study classifies and identifies the research methods employed in these articles. The process involves two main steps. Firstly, a suitable classification system of research methods is selected. Secondly, based on the classification system, a technique of automatic classification of research methods is used to identify the research methods of academic papers in the corpus and obtain the results of classification of research methods.

Classification system of research methods for academic papers in the field of LIS: Regarding the framework for research methods in the LIS field, the mainstream classification systems currently used in research primarily include two approaches. The first is the classification system proposed by Järvelin and Vakkari (1990). These scholars initially introduced a framework encompassing research strategies and methods, encoding data collection methods in academic papers from a methodological perspective. This system has been consistently updated in subsequent studies, though its core content remains largely unchanged (Järvelin & Vakkari, 1990; Järvelin & Vakkari, 1993; Järvelin & Vakkari, 2021). The second is the classification system proposed by Chu and Ke (2017), which focuses on data collection methods. By analyzing articles published in three prominent LIS journals—JASIST, LISR, and JOD—they developed a classification system comprising 16 data collection methods. Considering factors such as the granularity of the classification systems and their historical development, this study adopts the methodological framework proposed by Zhang et al. (2023) to identify research methods in the corpus of academic papers. The specific classification system is detailed in Table 4.

Table 4. Classification system of research methods in LIS discipline (Zhang et al., 2023).

No.	*Method*	*Definition*
1	Bibliometrics	Bibliometrics is a method used for collecting publication and citation data.
2	Content analysis	Content analysis refers to collecting data by conducting systematic examination of texts or other passages in the contexts of their use.
3	Delphi study	The Delphi method is generally used for collecting data with a questionnaire from a group of experts to address a research problem in order to reach consensus and make forecasts via several rounds of exchanges.
4	Ethnography/field study	Ethnography and field study share many characteristics in data collection. Both can be applied when collecting data using multiple techniques, such as observation and interview, in a natural setting where participants live or work.
5	Experiment	Experiment is an established method for collecting data by following a procedure to test what is studied in either a laboratory or field setting, corresponding to laboratory experiments and field experiments described in(Palvia et al., 2007) list of research methods.
6	Focus groups	As a research method, focus groups refer to data collection via discussion of a research problem between a moderator and a group of participants.
7	Historical method	Historical method refers to collecting data by examining, synthesizing, summarizing, and interpreting existing published and unpublished materials related to a historical research problem.
8	Interview	Interview is a data collection technique where individual participants are asked questions relating to a research problem.
9	Observation	Observation is a method for gathering data via carefully and attentively watching and making notes on the subject being studied.
10	Questionnaire	Questionnaire, often known as survey, is a technique for data collection using a predefined list of questions.
11	Research diary/journal	Research diary or journal is a technique used to gather data about events, activities, thoughts, reflections, or other aspects by an individual who keeps the diary over a period of time.
12	Theoretical approach	Theoretical approach, as a research method, is a technique for gathering data through conceptual analysis, theoretical examination, or similar activities.

13	Think aloud protocol	Think aloud protocol is a research method intended to collect data about participants' cognitive activities via the verbal reports of their thoughts, called think alouds, while taking part in an experiment or performing some task.
14	Transaction log analysis	Transaction log analysis, as a research method, gains momentum when computerized systems are used for information processing and access.
15	Webometrics	Webometrics is defined as bibliometrics in the web environment, where webpages and websites are generally regarded as publications; with inlinks (i.e., links a webpage or site receives) being considered as citations and outlinks (i.e., links a webpage or site makes to others) being considered as references.
16	Other methods	Research methods other than the 15 mentioned above.

Selection of the classification model for research methods in LIS academic papers: Previous studies have primarily relied on manual coding to identify research methods in academic papers, a process that is both time-consuming and labor-intensive, while also heavily dependent on expert knowledge (Chu & Ke, 2017; Järvelin & Vakkari, 1993). Given the substantial scale of the full-text corpus of LIS academic papers constructed in this study, an automated approach to research method classification is employed to identify the primary research methods at the document level for each paper. Inspired by the CogLTX model designed by Ding et al. (2020), Zhang et al. (2023) adapted this model for the task of research method classification, developing the CogFT (Cognize Full Text) model. This model demonstrates superior performance compared to traditional deep learning models based on pre-trained language models. Specifically, the CogFT model effectively extracts full-text features of academic papers while mitigating the noise introduced by irrelevant descriptions of research methods. Consequently, this study adopts the CogFT model for the task of document-level research method identification.

Since a single paper may employ multiple research methods, the total number of identified methods exceeds the number of academic papers. Using the CogFT model to automatically classify research methods in the full-text corpus, the study ultimately obtains the classification results. The final classification yielded 31,401 distinct methodological instances drawn from 26,677 articles. Notably, 3,074 articles were found to incorporate multiple research methods. As illustrated in Figure 2, the top five research methods used in the papers are bibliometrics, experiment, questionnaire, theoretical approach, and content analysis, collectively accounting for over 75% of the total methods identified.

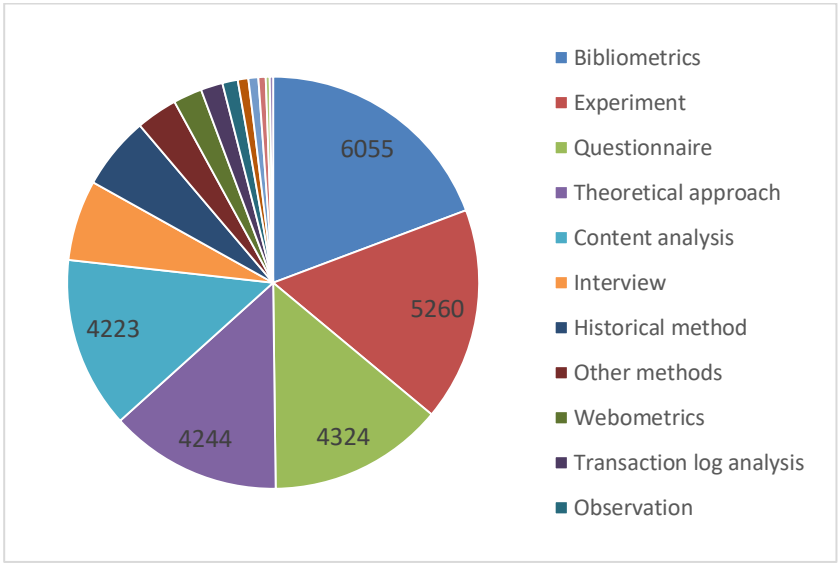


Figure 2. Classification results of research methods based on academic papers.

Data processing for scholars' academic careers in the LIS field

This study investigates the trajectory of research method selection among scholars in a specific field at different stages of their academic careers. In addition to the research methods identified earlier, it is necessary to perform author name disambiguation, calculate scholars' academic age. Based on these steps, we will select a subset of scholars to explore the trajectory of their methodological choices throughout their careers.

Scholar name disambiguation: To examine the trajectory of research method selection in scholars' academic careers, complete and accurate personal information is essential. This study utilizes OpenAlex [https://openalex.org/] to accomplish the task of author name disambiguation. OpenAlex is a free, open-access, large-scale scholarly resource indexing database that provides unique identifiers for various academic entities, including publications, authors, and institutions. It also offers multiple user-friendly API access methods. Among these, publication information can be retrieved using DOIs. Therefore, this study uses the DOIs from the metadata of academic papers to query OpenAlex, obtaining corresponding publication information and the unique identifiers of authors associated with each paper. These identifiers are then recorded and compared. Through this process, the study achieves accurate author name disambiguation results.

Calculation of academic age of scholars: To standardize the measurement of academic careers, this study defines a scholar's academic age as the time elapsed since their first publication. After completing the author name disambiguation process, the earliest publication of each author is retrieved from OpenAlex using their name. The publication year of this first paper is then extracted and used as the starting point for calculating academic age. Based on this starting point, the academic

age of a scholar at the time of publishing a subsequent paper is calculated by taking the difference between the publication year of the paper and the year of their first publication, then adding one. The formula for calculating academic age is as follows:

$$AAS = PYA - EPY + 1 \quad (1)$$

Where, AAS stands for Academic Age of Scholar, PYA stands for Publication Year of Article, and EPY signifies the Earliest Publication Year. It is important to note that a scholar's academic age does not necessarily correspond to a specific range in their actual chronological age, as the real age at which scholars publish their first paper may vary. Therefore, this study employs academic age as the metric for investigating the use of research methods throughout scholars' academic careers.

Criteria for selecting research method data in LIS scholars' academic careers: After the above processing steps, the author has obtained the data of scholars' papers. Next, we will select scholars and summarize the relevant data of the papers they published during their research careers.

First, to ensure the completeness and comprehensiveness of the data, this study considers both the first author and the corresponding author of each academic paper. In the corpus of academic papers used in this study, 14,856 articles have the same individual as the first author and corresponding author, while 8,471 articles have different individuals in these roles. Accordingly, when counting authors, this study considers both the first author and corresponding author for articles where these roles are distinct. For articles where the first author and corresponding author are the same, the author is counted as a single individual.

Second, to ensure the validity and reliability of the data, it is necessary to remove outliers in scholars' academic age. The Interquartile Range (IQR) method, which is based on the quantiles of the data, is effective in excluding extreme values and is not influenced by outliers. Therefore, this study employs the IQR method to identify and remove outliers in academic age. The academic age data of the scholars were first sorted from smallest to largest. Formula (2) calculates the inter - quartile range. Q1 represents the lower quartile, which is the value at the 25th percentile. Q3 represents the upper quartile, which is the value at the 75th percentile. Second, values in the academic - age data that are less than the lower limit or greater than the upper limit may be regarded as outliers. Formula (3) and Formula (4) calculate the upper and lower limits of the academic age respectively.

Finally, the calculated outliers of the authors' academic ages are eliminated, and a total of 14,622 authors' data are obtained.

$$IQR = Q3 - Q1 \quad (2)$$

$$Upper = Q3 + 1.5 * IQR \quad (3)$$

$$Lower = Q1 - 1.5 * IQR \quad (4)$$

As shown in Figure 3, the distribution of academic age of all authors is demonstrated. Among the authors, those with an academic age of 1 constitute the largest group, significantly outnumbering authors at other academic ages.

In this study, to analyze the relationship between authors' academic age and their use of research methods, a reasonable classification of academic age was established. Prior to categorizing scholars, to ensure the validity of academic age, the 95th percentile value of academic age was selected as the upper limit, setting the maximum academic age at 61 years. Building on prior research, authors were divided into three categories based on academic age. Authors with an academic age less than 7 were defined as young scholars. Those with an academic age between 7 and 14 were middle - aged scholars. Those with an academic age greater than 14 were senior scholars (Chowdhary et al., 2024).

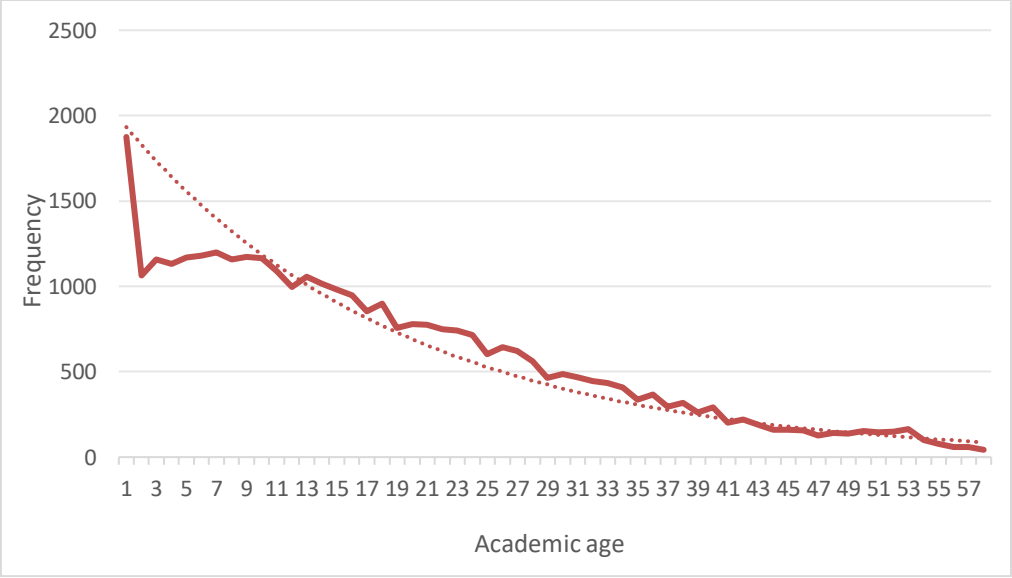


Figure 3. Distribution of authors' academic age.

Finally, the selection of scholars was conducted. Given the variability in the trajectory of research method selection across scholars' academic careers, this study focuses on scholars with longer and more active academic careers to capture the overall trends in methodological choices. Therefore, senior scholars with an academic age greater than 14 years and a consistent publication record at five-year intervals were selected. This resulted in a cohort of 435 senior scholars, encompassing 6,116 published articles.

Results

Correlation analysis of academic age and research methods of scholars in the field of LIS

In this section, we address ***RQ1*** by exploring whether differences exist in the selection of research methods among scholars at different academic ages. To achieve

this, statistical methods for difference analysis and testing are applied. Common methods for difference analysis include the t-test, analysis of variance (ANOVA), and the chi-square test. The chi-square test is suitable for scenarios where both independent and dependent variables are categorical. Therefore, this section employs the chi-square test to measure the frequency differences in method selection among scholars belonging to three distinct academic age groups. Each research method is independently subjected to a chi-square test. Since an article can only select a specific method once, the number of articles completed by scholars in different academic age groups serves as the basis for calculating expected frequencies. Figure 4 presents the results of the chi-square statistics.

Within the specific field, the usage proportions of different research methods exhibit significant variation. Among these, bibliometrics has the highest proportion at 30.02%, indicating that this method is the most commonly employed by scholars in the field. In contrast, focus groups, ethnography/field study, think aloud protocol, research diary/journal, and delphi study are used very infrequently, each accounting for less than 1% of the total. This suggests that these methods are rarely adopted in research. Out of the 16 research methods examined, only 6 show no significant differences in selection frequency across academic age groups. This indicates that scholars at different stages of their academic careers exhibit distinct preferences in their choice of research methods. When scholars are in the early stage of their academic careers, that is, when their academic age is less than 7, there are 3 methods they tend to choose. When scholars' academic age is between 7 and 14, there are 6 methods they prefer. When scholars' academic age is greater than 14, there are 4 methods they are inclined to select. Obviously, scholars in their middle - aged period tend to choose a larger variety of methods. In addition, this paper uses the chi - square value to judge the degree of significance of differences in method selection at different stages of the academic career. The top three methods with the largest chi-square values are webometrics (Chi2=114.8354***), bibliometrics (Chi2=43.3623***) and think aloud protocol (Chi2=11.5278**). Webometrics and bibliometrics are the methods preferred by academics in their younger and middle-aged years. Think aloud protocol is the method preferred by academics in their senior years.

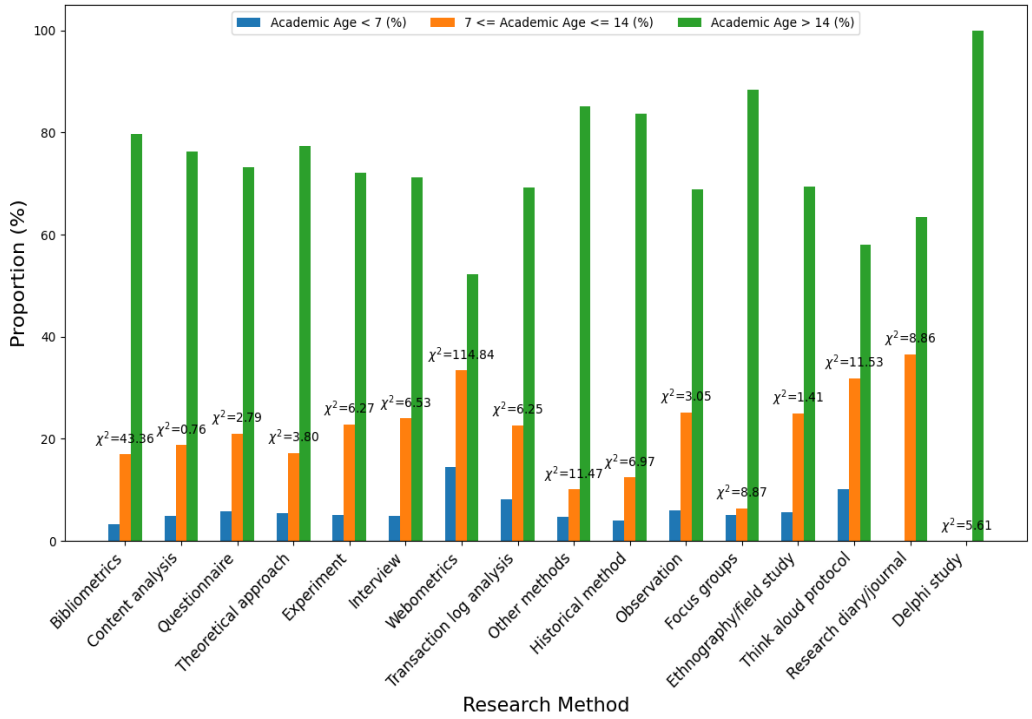


Figure 4. Statistical differences in the frequency of method selection among scholars in different academic age groups.

Differences in research methods used by scholars of different academic ages in different periods

To delve deeper into the variations in research method usage among scholars at different academic ages across various time periods, this study first examines the types of research methods employed by scholars at different career stages over time. It then focuses on the top five methods used by scholars in each academic age group and explores the evolving trends in the frequency of method usage based on publication years.

Types of research methods used by scholars in different academic age groups across publication periods: To investigate the diversity of research methods used by scholars at different career stages over time, this study constructs a heatmap based on five-year intervals of publication years and academic age groups. Since the number of publications varies across time periods, the data on the types of research methods used are normalized to ensure comparability.

As shown in Figure 5, the darker regions are predominantly concentrated in the period from 2000 to 2024 and among scholars with academic ages ranging from 1 to 50 years. This indicates that since 2000, scholars across various academic age groups have increasingly adopted a more diverse range of research methods. Furthermore, for each publication period after 2000, the number of research methods used initially increases and then decreases as scholars progress in their academic age.

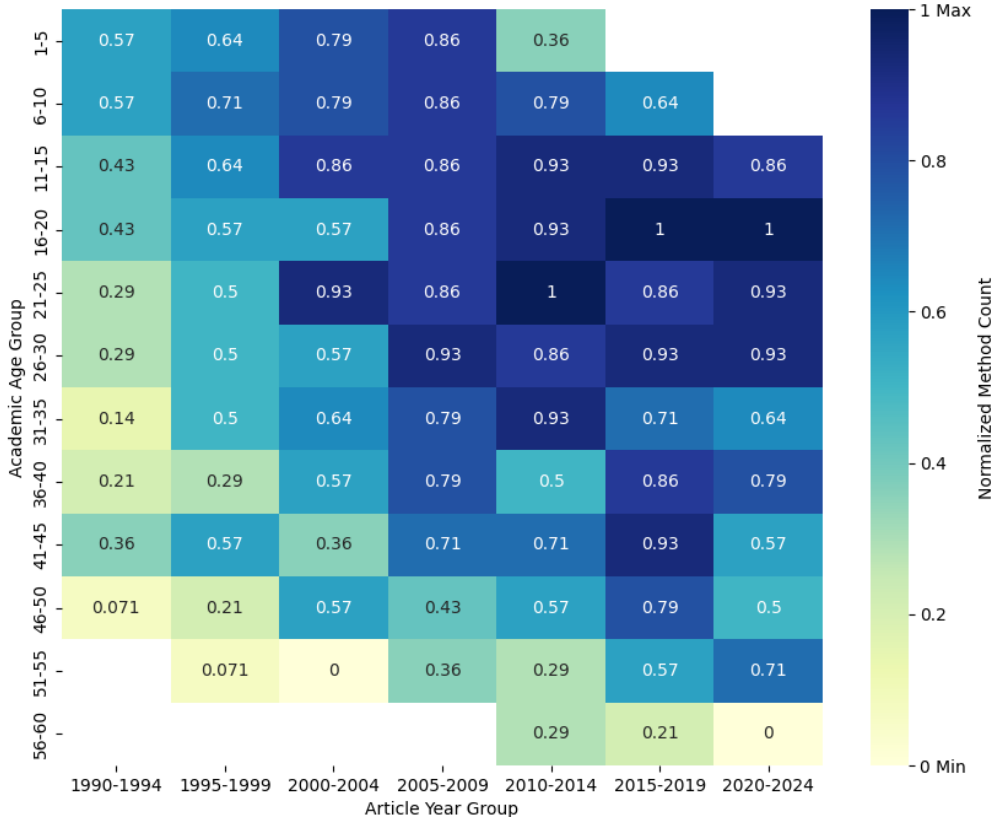


Figure 5. Heat map of the types of research methods used.

During the early period of 1990–1994, most academic age groups are represented by light green or light yellow hues. This suggests that, regardless of academic age, scholars during this time employed a relatively limited variety of research methods. From 2000 to 2014, the colors gradually deepen, particularly among scholars in the 11–45 academic age range, where the values reach as high as 0.93 or even 1. This indicates that scholars in this range utilized nearly all available types of research methods, reflecting a significant diversification in their methodological approaches. In the period of 2015–2024, the color distribution shifts again, with the hues for the 31–50 academic age group becoming lighter. The trend for the 11–30 academic age group shows that these scholars maintained a high diversity in research method usage over an extended period, likely due to their being in the prime of their academic careers, where they possess the capability and resources to experiment with a wide range of methodologies.For scholars in the 46+ academic age group, the overall number of research methods used is relatively low. This may be attributed to their methodological preferences having stabilized or to physical and other constraints limiting their ability to employ certain methods.

As shown in Figure 5, the diversity of research methods used exhibits dynamic changes across different academic age groups and publication periods. Over time, there is an overall trend toward increased methodological diversity, though the extent and timing of these changes vary among academic age groups. The middle - aged academic age group has maintained a high level of research method diversity over a

long period. Young scholars are continuously increasing the number of types of research methods they use, while senior scholars remain relatively stable.

Top five research methods used by scholars of different academic age groups: In order to deeply analyse which types of research methods are more popular among scholars at different stages of their academic careers, this paper summarizes the annual percentage of the top five research methods used by different academic age groups. The specific situation is shown in Figure 6.

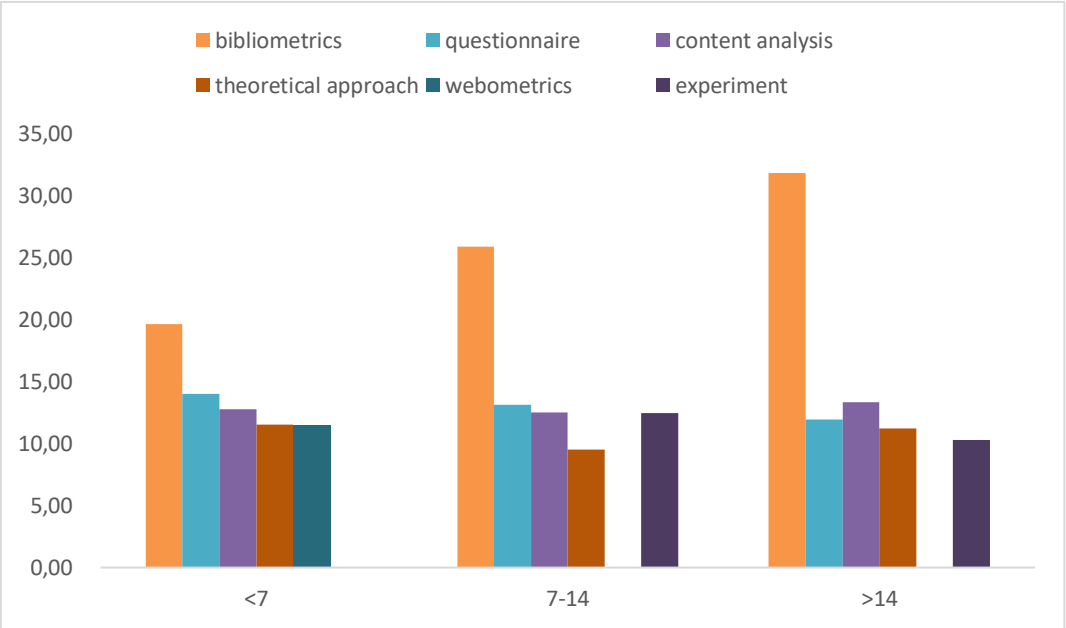


Figure 6. Top five research methods by usage proportion across different academic age groups.

As depicted in Figure 6, the most frequently used research methods among scholars remain relatively consistent across different academic age groups. For senior scholars, bibliometrics consistently ranks first in usage, with its proportion showing an upward trend. This indicates that bibliometric is highly favored by scholars, effectively aiding those engaged in long-term research endeavors with tasks such as literature analysis. It also underscores the dominant role of bibliometrics in the field of information science. Questionnaire and content analysis maintain stable usage proportions across all academic age groups, consistently ranking second and third, respectively. This reflects the broad applicability and enduring demand for these methods. Theoretical approach also persists throughout scholars' academic careers, highlighting the guiding role of theoretical research in academic inquiry. Webometrics ranks fifth in usage among younger scholars, indicating its popularity within this group. Meanwhile, experiment exhibits relatively stable usage proportions among mid-career and senior scholars, ranking fourth and fifth, respectively. This suggests that experiment becomes an important research tool as scholars accumulate experience and enhance their research capabilities.

The usage proportions of research methods among scholars in different academic age groups also vary over time. Specific details are illustrated in Figure 7.

As shown in Figure 7, the trend of the top five research methods in terms of percentage of use varies slightly across different academic ages in different periods of time. Bibliometrics covers the range of academic careers of scholars in all periods of time and is consistently high in terms of percentage of use. It is followed by content analysis, experiment and questionnaire. This confirms the trend of the overall top five used research methods as reflected in Figure 6. Since 2000, webometrics has been highly favored by young scholars, and it ranked fifth among the methods used by middle - aged scholars from 2005 to 2009. This may be attributed to the fact that young scholars from 2000–2004, as they advanced in age and experience, transitioning into middle - aged scholars, retained their preference for bibliometrics. For theoretical approach, the method was highly preferred by scholars at all academic career stages from 1990-2000, with a share of around 20%. However, its ranking gradually declined after 2000 and disappeared from the top five list after 2015. This shift may be linked to the rise of emerging technologies, such as machine learning models, which have increasingly been applied in academic papers, potentially displacing other traditional methods. Certain methods, such as transaction log analysis and focus groups, appear prominently only in specific periods and academic age groups. This may reflect the methodological preferences of particular scholars during those times.

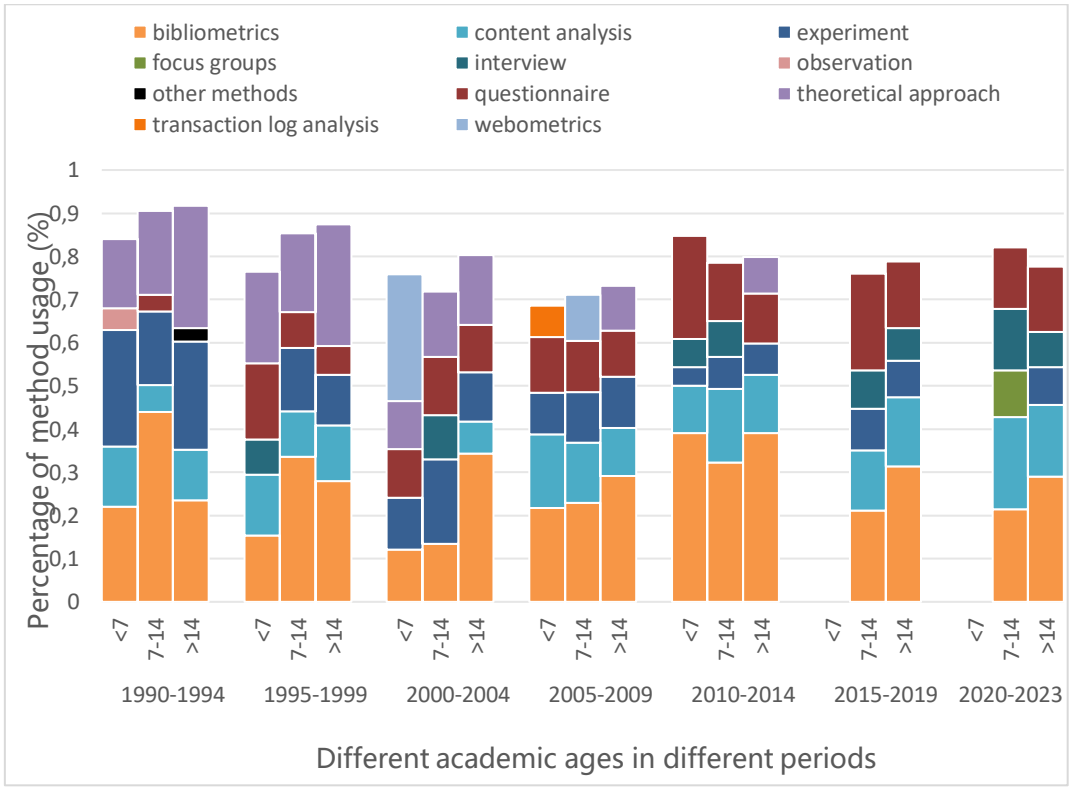


Figure 7. Top five research methods used by different academic age groups.

Overall, scholars in different academic age groups exhibit variations in their use of research methods across different time periods. Over time, the usage proportions of certain methods, such as bibliometrics and content analysis, have gradually increased across all academic age groups. In contrast, the usage proportions of more traditional methods, such as interview and theoretical approach, have declined. These shifts reflect broader trends in academic research and the influence of technological advancements on methodological preferences.

Evolution of research method usage among scholars at different career stages: Figure 8 presents the evolving trends in the frequency of usage for the 16 research methods among scholars in different academic age groups. Overall, the usage frequency of most methods shows significant fluctuations between 1990 and 2020. This indicates that scholars' adoption of research methods has not been stable over the years, likely influenced by factors such as shifts in research hotspots, technological developments, and interdisciplinary integration. These fluctuations underscore the diversity and dynamism of research methodologies in academic inquiry. Moreover, for each research method, the trends in usage frequency appear consistent across the three academic career stages. This may be attributed to the inherent characteristics of the methods themselves, where a method gaining popularity in a particular period leads to its widespread adoption by scholars across all age groups.



Figure 8. Evolutionary trends in the use of different research methods by scholars at different stages of their academic careers.

In papers published by scholars in the senior stage of their careers, the use of methods such as bibliometrics, content analysis, interview, and questionnaire exhibits a pronounced upward trend. Notably, bibliometrics, which had relatively low usage frequency from 1990 to 1995, experienced rapid growth starting in 1995 and maintained high usage frequency between 2010 and 2020. This trend may be linked to the rapid development of scientometrics and the increasing emphasis on literature analysis in academia. In contrast, the use of experiment and theoretical approach remains relatively stable, indicating that theoretical research continues to hold a significant position in academic inquiry. During the period of 2005–2010, methods such as experiment, historical method, interview, observation, and transaction log analysis reached a notable peak in usage. This suggests that scholars during this five-year period were inclined to employ a diverse range of research methods rather than limiting themselves to commonly used or popular approaches.

Apart from the aforementioned methods, most other methods do not exhibit significant trends in usage frequency due to their inherently low adoption rates. When scholars are in the early stages of their careers, the use of webometrics shows a leading trend. This can be attributed to the influence of internet technology on academic research methods, as well as the greater willingness of younger scholars to adopt and apply emerging technologies. When scholars are in their middle age, the frequency of using all kinds of research methods increases compared with that in their younger age. This may be because scholars' careers are relatively stable in middle age and the valuation risk is relatively reduced. Therefore, scholars will try to use a variety of research methods to achieve self - breakthroughs and enhance their academic influence.

As shown in Figure 8, the trends in the usage frequency of different research methods between 1990 and 2020 vary significantly. Emerging methods, such as webometrics, exhibit rapid growth trends driven by technological advancements. In contrast, more traditional methods, such as questionnaire and theoretical approach, maintain relatively stable usage frequencies. The adoption of research methods is influenced by a variety of factors, including disciplinary developments, technological progress, and shifts in research hotspots. Scholars adapt their methodological choices over time to align with the practical demands of their research.

Trajectory of research method usage in LIS scholars' academic careers

In this section, we address ***RQ2*** by exploring the trajectory of research method usage in the academic careers of LIS scholars, both at the aggregate and individual levels. Building on the earlier analysis of the types of research methods used by scholars during their careers, this study examines the overall differences in method usage from 1990 to 2023. However, specific trends in the trajectory may be obscured by factors such as the popularity of certain methods. Therefore, this subsection focuses on scholars who published their first paper between 1970 and 1979. This cohort was selected to minimize the generational effects of academic age differences on method usage and because scholars in this decade produced a higher volume of publications compared to other ten-year intervals, making them particularly valuable for analysis.

Aggregate trajectory of research method usage in LIS scholars' academic careers: Figure 9 illustrates the evolving trends in research method usage among scholars who published their first paper between 1970 and 1979, as their academic age increased. Methods such as bibliometrics, content analysis, experiment, questionnaire and theoretical approach were widely used across different academic ages. Notably, experiment was more frequently employed when scholars were between 13 and 28 academic years old, while questionnaire became more prevalent after scholars reached 29 academic years of age. Over the course of their academic careers, scholars exhibited a trend toward greater diversity in the types of research methods they used as they aged.

To better demonstrate this relationship, we created an interactive heatmap [https://jiayihao - njust.github.io/tra/]. This interactive graph collected data from the group of scholars whose earliest publication time was from 1970 to 1979. It can dynamically display the changes in the research methods used by scholars each year as their academic age increases. At the bottom of the interactive graph, there is a "Pause" button, which allows users to pause at any time to view the usage trajectories of research methods in the academic careers of scholars in the LIS field in any specific year from 1990 to 2023. The detailed information of the selected scholars for this graph can be found in Table A of Appendix.

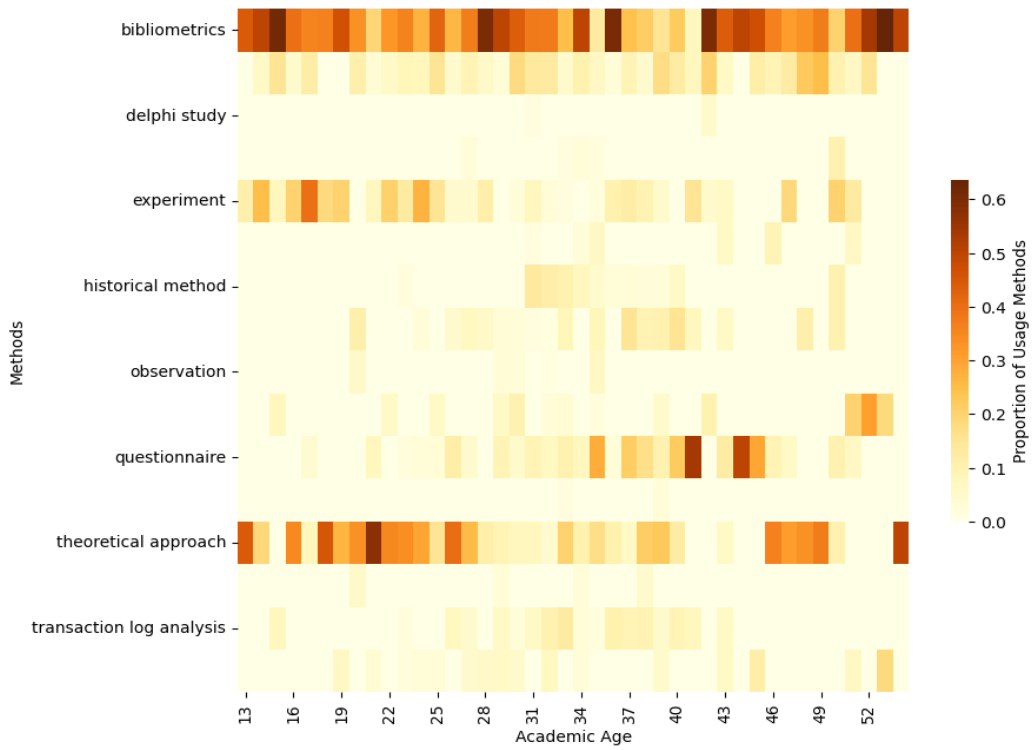


Figure 9. Evolution of research method usage among scholars at different career stages.

Individual trajectories of research method usage in scholars' academic careers: To further explore the characteristics of research method usage in scholars' academic careers, this study randomly selects four senior scholars and conducts a detailed analysis of their methodological trajectories. Due to limited data availability, the analysis of these scholars' careers is based on their publications in the 15 selected journals between 1990 and 2023.For each scholar, the analysis focuses on the following aspects:the most frequently used research methods, the combination of methods employed, and the trends in changes to their research method usage over time.

Mike Thelwall is a male scholar whose first publication appeared in 2000. Over the course of his academic career, he has employed eight research methods, with the most frequently used being bibliometrics, webometrics, and content analysis. He has also utilized combined methods in his research, primarily pairing commonly used methods. During his early-career stage, he predominantly relied on webometrics. As he transitioned into the mid-career stage, his methodological repertoire expanded to include webometrics and content analysis, and he began incorporating combined methods into his research. In his senior-career stage, his most frequently used methods were bibliometrics, webometrics, and content analysis, with an increased reliance on combined methods. This demonstrates that his selection and use of research methods evolved in stages as he advanced in age and experience.

Amanda Spink is a female scholar whose first publication appeared in 1992. Throughout her academic career, she has employed ten research methods, with the most frequently used being transaction log analysis, questionnaire, content analysis, experiment, and theoretical approach. She has also employed combined methods in her research, including combinations of commonly used methods as well as pairings of common and less common methods. In some publications, she used up to four combined methods. Notably, the diversity of methods she employed remained consistent across different stages of her academic career, indicating her proficiency and habitual use of various methodologies to support her research endeavors.

Noa Aharony is a female scholar whose first publication appeared in 2006. Over her academic career, she has employed four research methods, with the most frequently used being questionnaire and content analysis. She has also utilized combined methods in her research, pairing commonly used methods with less common ones. During her early-career stage, her primary methods were questionnaire and content analysis. As she transitioned into the mid-career stage, the use of questionnaire increased significantly, while the use of content analysis declined relatively. She began employing combined methods and other methodologies during this period. In her senior-career stage, her most frequently used method was questionnaire. This suggests that, while her methodological choices exhibited a brief period of diversification during her mid-career stage, they ultimately stabilized. This stability may be attributed to the constraints of her research topics or her habitual preferences in method selection.

José Ortega is a male scholar whose first publication appeared in 2003. Throughout his academic career, he has employed four research methods, with the most frequently used being webometrics, content analysis, and bibliometrics. He has also

utilized combined methods in his research, primarily pairing commonly used methods. During his early-career stage, his sole research method was webometrics. In his mid-career stage, his methodological choices evolved from webometrics to content analysis, then to a combination of content analysis and bibliometrics, and finally back to bibliometrics. In his senior-career stage, his most frequently used methods were content analysis and bibliometrics. This indicates a notable trend of methodological diversification during his mid-career stage.

From the trajectories of research method usage among the four scholars described above, it is evident that during the mid-career stage, scholars exhibit a tendency to employ a diverse range of research methods, accompanied by an increase in publication output. The most frequently used research methods shift as scholars advance in age and experience, likely influenced by the popularity of certain methods and research topics during different periods. Throughout their academic careers, scholars experiment with various combinations of research methods, whether pairing commonly used methods or combining less common methods with popular ones. This reflects their flexibility and adaptability in applying research methodologies to their work.

Discussion

Research implications

Theoretical implications: The basic aim of this study was to explore the trajectory of research methodology use in scholars' academic careers based on journal articles in the field of LIS. Additionally, this study makes two unique contributions to the understanding of research method usage.

First, we combine the automatic categorization of research methods with scholars' academic careers to explore the relationship between scholars' academic age and research use. From the perspectives of cognitive and sociological theories, scholars of different academic ages may have different personal cognitive understandings that affect their choice and use of research methods. Scholars of different academic ages also have different preferences in the use of research methods.

Second, this study provides a comprehensive and dynamic overview of research method usage among LIS scholars from 1990 to 2023. It highlights innovative directions and the application of cutting-edge research methods within the LIS field, offering theoretical insights and guidance for disciplinary development and innovation.

Practical Implications: From the perspective of individual scholars, by examining the differences in research method usage among scholars of different academic ages, scholars can learn from the methodological trajectories of senior scholars with similar backgrounds or research interests. Scholars at different stages of their academic careers may choose different research methods based on their evolving research interests and the contextual demands of their time. On a personal level, paying attention to the research topics and methods favored by scholars of different academic ages can help uncover hidden patterns in the relationship between academic age and methodological choices. Additionally, young scholars can learn

from the use of research methods by senior scholars, thus enriching the variety of research methods used in their own academic research and promoting their personal career development.

From an institutional perspective, this study offers recommendations for developing academic guidance programs that promote methodological diversity. The findings reveal that scholars of different age groups exhibit distinct preferences for research methods, with certain methods gaining varying levels of popularity across academic age groups. However, the LIS field is characterized by methodological diversity, and scholars at different academic ages may exhibit varying degrees of methodological specialization, sometimes leading to a narrow focus on specific methods. This study enhances understanding of such dynamics and provides insights for institutions to design academic guidance programs that encourage methodological diversity and innovation.

Research limitations

The study of research method usage trajectories in the academic careers of LIS scholars still faces several challenges. First, the scope of this study is limited, as it only includes data from 14 LIS journals published between 1990 and 2023. Future research aims to expand the data sources to encompass the complete publication records of scholars throughout their academic careers. Second, while this study collected and visualized data on research method usage in scholars' academic careers, it did not delve into the underlying reasons for their methodological choices. Structural factors—such as funding dynamics, the influence of journal policies, and broader disciplinary trends—remain underexplored. In subsequent work, additional factors such as research topics, scholar gender, and country of origin will be incorporated to explore the influences on scholars' selection and use of research methods. Finally, due to time constraints, this study analyzed the methodological trajectories of only a subset of scholars. Future research will consider including a larger cohort of senior scholars in the field to comprehensively explore research method usage trajectories and derive a paradigm for methodological practices in scholars' academic careers.

Conclusions and future research

We draw on data from 14 authoritative journals in the LIS field published between 1990 and 2023, selecting a subset of scholars to explore the trajectory of research method usage in their academic careers.

Based on the results, several conclusions can be drawn about the two research questions posed in this study. We found that the research methods commonly used by scholars in the field of LIS will change with the growth of age and seniority, which may be affected by factors such as popular research methods and research topics at different times. Over the course of their academic careers, scholars exhibit an initial increase followed by a decline in the diversity of research methods used. They also demonstrate a tendency to combine multiple methods, whether pairing commonly used methods or integrating less common methods with popular ones, reflecting their flexibility and adaptability in applying research methodologies.

Scholars' use of research methods is influenced by a variety of factors, including disciplinary developments, technological advancements, and shifts in research hotspots. As a result, scholars adapt their methodological choices over time to align with the practical demands of their research.

In future work, we intend to incorporate information such as research topics, genders, and research backgrounds into the study. Building upon initial findings from chi-square tests, which reveal statistically significant variations in methodological preferences across career stages, we will employ more advanced analytical techniques to identify causal mechanisms underlying these patterns. In addition, we would like to expand the data sources, starting from individual scholars, to obtain the papers published by scholars during their academic careers that cover a wider range of journals.

Acknowledgments

This study has received support from the National Natural Science Foundation of China (Grant No.72074113).

References

Abramo, G., D'Angelo, C. A., & Murgia, G. (2016). The combined effects of age and seniority on research performance of full professors. Science and Public Policy, 43(3), 301–319.

Ao, W., Lyu, D., Ruan, X., Li, J., & Cheng, Y. (2023). Scientific creativity patterns in scholars' academic careers: Evidence from PubMed. Journal of Informetrics, 17(4), 101463.

Aref, S., Zagheni, E., & West, J. (2019, November). The demography of the peripatetic researcher: Evidence on highly mobile scholars from the Web of Science. In International conference on social informatics (pp. 50-65). Cham: Springer International Publishing.

Azoulay, P., Fons-Rosen, C., & Zivin, J. S. G. (2019). Does Science Advance One Funeral at a Time? American Economic Review, 109(8), 2889–2920.

Badar, K., M. Hite, J., & F. Badir, Y. (2014). The moderating roles of academic age and institutional sector on the relationship between co-authorship network centrality and academic research performance. Aslib Journal of Information Management, 66(1), 38–53.

Bu, Y., Murray, D. S., Xu, J., Ding, Y., Ai, P., Shen, J., & Yang, F. (2018). Analyzing scientific collaboration with "giants" based on the milestones of career. Proceedings of the Association for Information Science and Technology, 55(1), 29–38.

Chan, H. F., & Torgler, B. (2020). Gender differences in performance of top cited scientists by field and country. Scientometrics, 125(3), 2421–2447.

Chowdhary, S., Gallo, L., Musciotto, F., & Battiston, F. (2024). Team careers in science: formation, composition and success of persistent collaborations. arXiv preprint arXiv:2407.09326.

Chu, H. (2015). Research methods in library and information science: A content analysis. Library & Information Science Research, 37(1), 36–41.

Coomes, O. T., Moore, T., Paterson, J., Breau, S., Ross, N. A., & Roulet, N. (2013). Academic Performance Indicators for Departments of Geography in the United States and Canada. The Professional Geographer, 65(3), 433–450.

Costas, R., Nane, G. F., & Lariviere, V. (2015). Is the Year of First Publication a Good Proxy of Scholars Academic Age?. In International Conference on Scientometrics &

Informetrics (pp. 988-998). Retrieved from https://www.issi-society.org/proceedings/issi_2015/0988.pdf
Cui, H., Wu, L., & Evans, J. A. (2022). Aging scientists and slowed advance. arXiv preprint arXiv:2202.04044.
Ding, M., Zhou, C., Yang, H., & Tang, J. (2020). Cogltx: Applying bert to long texts. Advances in Neural Information Processing Systems, 33, 12792-12804.
Győrffy, B., Csuka, G., Herman, P., & Török, Á. (2020). Is there a golden age in publication activity?—An analysis of age-related scholarly performance across all scientific disciplines. Scientometrics, 124(2), 1081–1097.
Hayman, R. & Smith, E. (2020). Mixed Methods Research in Library and Information Science: A Methodological Review. Evidence Based Library and Information Practice, 15(1), 106–125.
Heting Chu & Qing Ke. (2017). Research methods: What's in the name? Library & Information Science Research, 39(4), 284–294.
Järvelin, K., & Vakkari, P. (1990). Content Analysis of Research Articles in Library and Information Science. Library & Information Science Research, 12, 395-421.
Järvelin, K., & Vakkari, P. (1993). The evolution of library and information science 1965–1985: A content analysis of journal articles. Information Processing & Management, 29(1), 129–144.
Järvelin, K., & Vakkari, P. (2021). LIS research across 50 years: Content analysis of journal articles. Journal of Documentation, 78(7), 65–88.
Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research interest evolution. Nature Human Behaviour, 1(4), 0078.
Kumar, S., & Ratnavelu, K. (2016). Perceptions of Scholars in the Field of Economics on Co-Authorship Associations: Evidence from an International Survey. PLOS ONE, 11(6), e0157633.
Liang, G., Hou, H., Ding, Y., & Hu, Z. (2020). Knowledge recency to the birth of Nobel Prize-winning articles: Gender, career stage, and country. Journal of Informetrics, 14(3), 101053.
Liao, C. H. (2017). Reopening the Black Box of Career Age and Research Performance. In J. Zhou & G. Salvendy (Eds.), Human Aspects of IT for the Aged Population. Applications, Services and Contexts (Vol. 10298, pp. 516–525). Springer International Publishing.
Lou, W., Su, Z., He, J., & Li, K. (2021). A temporally dynamic examination of research method usage in the Chinese library and information science community. Information Processing & Management, 58(5), 102686.
Lund, B. D., & Wang, T. (2021). An analysis of research methods utilized in five top, practitioner-oriented LIS journals from 1980 to 2019. Journal of Documentation, 77(5), 1196–1208.
Milojević, S. (2012). How Are Academic Age, Productivity and Collaboration Related to Citing Behavior of Researchers? PLoS ONE, 7(11), e49176.
Nane, G. F., Larivière, V., & Costas, R. (2017). Predicting the age of researchers using bibliometric data. Journal of Informetrics, 11(3), 713–729.
Packalen, M., & Bhattacharya, J. (2019). Age and the Trying Out of New Ideas. Journal of Human Capital, 13(2), 341–373.
Palvia, P., Pinjani, P., & Sibley, E. H. (2007). A profile of information systems research published in Information & Management. Information & Management, 44(1), 1–11.

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2015). Within- and between-department variability in individual productivity: The case of economics. Scientometrics, 102(2), 1497–1520.

Robinson-Garcia, N., Costas, R., Sugimoto, C. R., Larivière, V., & Nane, G. F. (2020). Task specialization across research careers. eLife, 9, e60586.

Simoes, N., & Crespo, N. (2020). A flexible approach for measuring author-level publishing performance. Scientometrics, 122(1), 331–355.

Sugimoto, C., Sugimoto, T., Tsou, A., Milojevic, S., & Larivière, V. (2016). Age stratification and cohort effects in scholarly communication: A study of social sciences: Scientometrics, 109.

van den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: A longitudinal case study. Scientometrics, 106, 143–162.

Wang, W., Yu, S., Bekele, T. M., Kong, X., & Xia, F. (2017). Scientific collaboration patterns vary with scholars' academic ages. Scientometrics, 112(1), 329–343.

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. Nature Communications, 10(1), 3439.

Zhang, C., Tian, L., & Chu, H. (2023). Usage frequency and application variety of research methods in library and information science: Continuous investigation from 1991 to 2021. Information Processing & Management, 60(6), 103507.

Zhang, L., Qi, F., Sivertsen, G., Liang, L., & Campbell, D. (2024). Gender differences in the patterns and consequences of changing research directions in scientific careers. Quantitative Science Studies, 5(4), 882–905.

Appendix

Table A. Information on scholars with first publications between 1970 and 1979.

Earliest pub year	*Author name*	*Number of publications (1990-2023)*
1970	E. Michael Keen	10
1970	J. A. García	49
1970	Jaime A. Teixeira da Silva	12
1970	V.K. Singh	21
1970	W. W. Hood	14
1971	Anthony F. J. van Raan	25
1971	Barrie Gunter	12
1971	David Nicholas	55
1971	Michael E. D. Koenig	12
1971	Peter Vinkler	26
1972	Donald O. Case	10
1972	Peter Hernon	14
1973	Henry Small	12
1973	Ian Ruthven	12
1973	Jennifer Rowley	56
1973	M. H. Heine	10
1973	Peter Williams	16
1974	Mingyang Wang	11
1975	Gangan Prathap	17
1976	G.E. Gorman	56
1976	Maria Pinto	76
1976	R. Rada	16
1977	Birger Hjørland	29
1977	Howard D. White	34
1977	Hsin Hsin Chang	11
1977	Mark E. Rorvig	20
1978	Blaise Cronin	36
1978	Jin Zhang	34
1978	Leo Egghe	226
1978	Peter Willett	27
1979	Jin Ha Lee	13
1979	Nigel Ford	23
1979	Philip M. Davis	16

https://doi.org/10.51408/issi2025_050

Transforming Researcher Evaluation: A New Global Platform to Measure Impact Across Disciplines

Balázs Győrffy[1], Boglárka Weltz[2], István Szabó[3]

[1] *gyorffy.balazs@yahoo.com*
Department of Bioinformatics, Semmelweis University, Tuzolto u. 7-9, H-1094, Budapest (Hungary)
Dept. of Biophysics, University of Pecs, Szigeti u. 12, H-7624, Pecs (Hungary)
HUN-REN Research Centre for Natural Sciences, Magyar tudosok korutja 2, H-1117, Budapest (Hungary)

[2] *weltz.boglarka@gmail.com*
Department of Bioinformatics, Semmelweis University, Tuzolto u. 7-9, H-1094, Budapest (Hungary)

[3] *istvan.szabo.phd@gmail.com*
Óbuda University, Bécsi út 96/B, H-1034, Budapest (Hungary)

Abstract

Ranking researchers based solely on raw metrics such as citation counts or the H-index can introduce significant biases. These measures often disadvantage early-career scientists and those working in disciplines with distinct publication norms. To address this inequity, we aimed to create a global, field-adjusted reference for evaluating scientific productivity.

We developed a comprehensive worldwide reference database encompassing 19 scientific disciplines. Using data from Scopus, we analyzed the most recent 5,000 researchers across 174 sub-fields. To account for disciplinary differences, we incorporated diverse publication types into the analysis tailored to each domain.

Our reference dataset includes 507,233 researchers from across the globe and facilitates the calculation of expected values for H-index, annual citations, and recent publications (within the past five years) for each percentile in every discipline. These benchmarks were stratified by career stage, assessed at each year after a researcher's first publication. A composite score was developed to rank publication performance into deciles (D1–D10), where D1 represents the highest level of achievement. Importantly, only data from researchers within the same career stage and scientific domain are used for comparison, ensuring fair and context-sensitive evaluations. To enhance accessibility, we established a web portal (www.scientometrics.org) to facilitate researcher benchmarking.

This age- and discipline-normalized international database promotes the application of responsible metrics, offering a robust framework for global scientometric rankings. By providing an online analysis platform, we enable researchers, institutions, and policymakers to determine expected levels of scholarly output at the individual level while fostering fairness and equity in academic evaluation.

Introduction

Researchers' performance is often evaluated using quantitative metrics such as publication counts, citations, and the H-index, which provide an initial overview of achievements. These metrics are widely used, driven by the "Publish or Perish" culture that prioritizes publication volume. Tools like Web of Science, Scopus, and Google Scholar supply these indicators, serving grant agencies, academic

committees, and university rankings (Szluka et al., 2023; Győrffy et al., 2023). However, exclusive reliance on these metrics introduces three key limitations.

First, they embed systemic biases. Researchers with longer careers naturally accumulate higher values, disadvantaging early-career scientists. For instance, peak productivity in fields like economics is often reached eleven years post-PhD (Lan et al., 2023). Metrics also vary across disciplines; an exceptional H-index in one field may be average in another (Győrffy, Csuka et al., 2020).

Second, academic productivity often plateaus and declines after a researcher's "Golden Age" (Győrffy, Csuka et al., 2020; Alchokr et al., 2022), and traditional metrics fail to account for those non-active "giants" whose past influential work inflates their indicators, skewing evaluations of their current relevance.

Third, authorship conventions complicate contributions. In many fields, first authorship signifies significant involvement, while the last author often represents the supervising researcher. Middle authorship contributions vary widely. Though these conventions are less established in the arts and humanities, they are gaining acceptance.

Evaluating scientific impact using single metrics like citation counts or H-index provides an oversimplified view of scholarly contributions. A study of over 84,000 scientists revealed that traditional metrics fail to capture the complexities of modern research, especially in fields with extensive multi-authorship (Ioannidis et al., 2016). Notably, only 322 of the top 1,000 scientists ranked by comprehensive metrics appear in the top 1,000 based on total citations, and some highly-cited researchers have never been first, last, or sole authors. These discrepancies, along with significant disciplinary differences in publication patterns, highlight the need for a multi-parameter approach to measuring productivity.

Our prior analyses have identified robust predictors of future scientific output. A large-scale evaluation of grant allocation in Hungary, analyzing 42,905 review reports for 13,303 proposals, found that H-index, yearly independent citation counts, and publications in top-tier journals (Q1) were the strongest predictors of future success, dramatically outperforming reviewer-assigned scores (Győrffy, Herman et al., 2020). Similarly, an analysis of Hungarian Momentum grant recipients showed that total citations, H-index, and publication impact factors strongly correlated with future productivity, while factors like gender, degree, or international grants showed no significant effect (Győrffy et al., 2018, Tóth et al. 2024).

Building on these findings, we developed a novel evaluation system that incorporates validated metrics while normalizing for age, discipline, and authorship position. Initially piloted for Hungarian researchers, where it achieved high national engagement (Győrffy et al., 2022), we redesigned the system for global application using Scopus data. The updated framework addresses data source differences and categorization standards, enabling accurate and equitable global assessment. Overall, here we introduce a fairer framework for researcher evaluation, mitigating biases inherent in traditional methods. Our online platform now provides a user-friendly tool for assessing and comparing approximately 20 million researchers across diverse scientific disciplines.

Methods

Data source

We used Elsevier's Science-Metrix database to classify fields, grouping 174 subfields into 19 broader fields across applied sciences, arts, humanities, health sciences, economics, and natural sciences. One field, visual and performing arts, was excluded due to insufficient data.

Publication and citation data were obtained from Scopus using its Search and Citation Overview APIs. Authors with at least five publications were included, prioritizing the most recent 5,000 per subfield for manageability. Independent citations were extracted for all articles.

Scientometric parameters

We computed H-index, yearly independent citation counts, and the number of publications in the past five years (Q1 journals preferred). Age was defined as years since the first publication. High-impact publications were identified using citation thresholds based on Web of Science's top 1% thresholds (**Table 1**).

Table 1. The research subfields were grouped into broader research fields, representing the five major domains of science. The "Scopus author n" column refers to the number of researchers from Scopus who were included in the reference database. The "Citation/year" column indicates the annual independent citation count required to classify a publication as high-impact within the specified scientific field.

Domain	Field	Scopus author n	Citation / Year	Source
Applied Sciences	Agriculture, Fisheries & Forestry	25487	14.3	WoS
	Built Environment & Design	6558	16.2	Computed
	Enabling & Strategic Technologies	18997	16.2	Computed
	Engineering	38551	16.7	WoS
	Information & Communication Technologies	24924	17.6	WoS
Arts & Humanities	Communication & Textual Studies	5824	11.5	= Social sciences
	Historical Studies	16447	11.5	= Social sciences
	Philosophy & Theology	10083	11.5	= Social sciences
Economic & Social Sciences	Economics & Business	32987	16.6	WoS
	Social Sciences	38186	11.5	WoS
Health Sciences	Biomedical Research	40078	32.6	WoS
	Clinical Medicine	101402	18.1	WoS
	Psychology & Cognitive Sciences	18823	16.0	WoS
	Public Health & Health Services	20315	22.2	Computed
Natural Sciences	Biology	19246	21.0	WoS
	Chemistry	21563	21.2	WoS
	Earth & Environmental Sciences	20612	16.3	WoS
	Mathematics & Statistics	13743	7.2	WoS
	Physics & Astronomy	33397	16.2	WoS

Discipline-specific adjustments

Fields were grouped into three categories to account for differences in publication practices. For most sciences, only first- or last-authored Q1 publications were considered, while arts, humanities, and social sciences included all Scopus-indexed outputs (**Figure 1**).

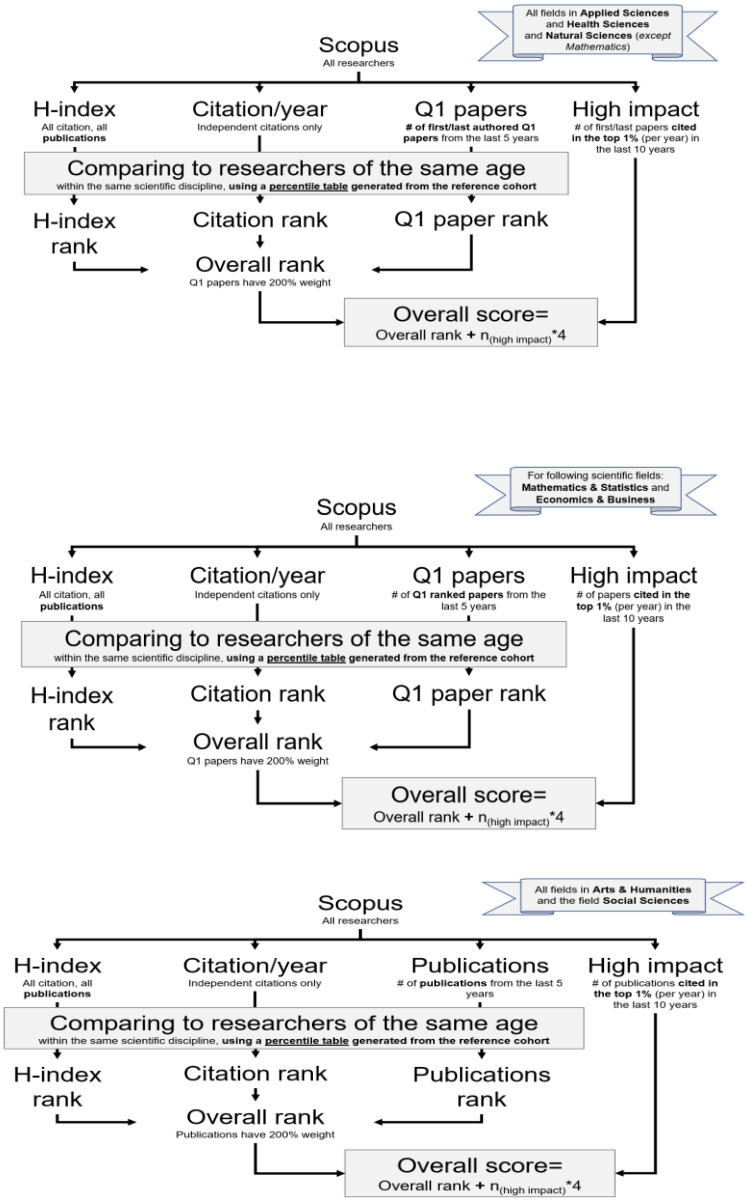


Figure 1. The overall score is calculated using distinct pipelines tailored to each major scientific discipline in order to account for the unique publication patterns specific to each field. One key aspect of this analysis involves the normalization of publication age relative to the first scientific publication. For example, the H-index of a researcher who has been active for 10 years is compared to the H-index at the same career stage for all researchers who have been active for more than 10 years.

Percentile tables and ranking

We generated percentile tables for H-index, citations, and publications by career age and field. Scores were averaged, with publication output weighted double. High-impact publications increased scores by a fixed value. Researchers were ranked into deciles (D1–D10) based on their total score.

An online portal, built using R Shiny, computes rankings for Scopus-indexed researchers. Users can input a name or Scopus ID to retrieve and rank data instantly. The platform is accessible at www.scientometrics.org/scopus.

Results

Database overview

For establishing the reference cohort, we analyzed 507,223 researchers across 174 subfields using Scopus data, excluding fields with insufficient publications (e.g., Education, Music, Law). China and the U.S. were the top contributors, each exceeding 86,000 researchers. Publication years peaked between 2015–2019, with fewer recent authors meeting the five-publication threshold. Researchers in natural sciences dominated publication counts, while arts & humanities had the least.

Scientometric analysis

We calculated the H-index, annual independent citations, and publication counts for each researcher at each career age with yearly bins. Independent citations were used for yearly citation metrics, while H-index calculations included all citations, aligning with Hirsch's original definition. Discipline-specific publication patterns were considered, ensuring accurate cross-field comparisons (**Figure 1**).

To speed up analysis, percentile thresholds for H-index, yearly citations, and publication counts at each years post-first publication were established across 19 disciplines. These thresholds allow rapid researcher assessment and are available for download on the scientometrics.org website.

Online analysis portal

We developed an online platform for researcher evaluation. Users can input a researcher’s name and discipline to retrieve Scopus data and generate rankings based on H-index, citations, and publication counts. Visualizations include percentile-based distributions across disciplines, offering detailed insights into a researcher's relative performance (**Figure 2**).

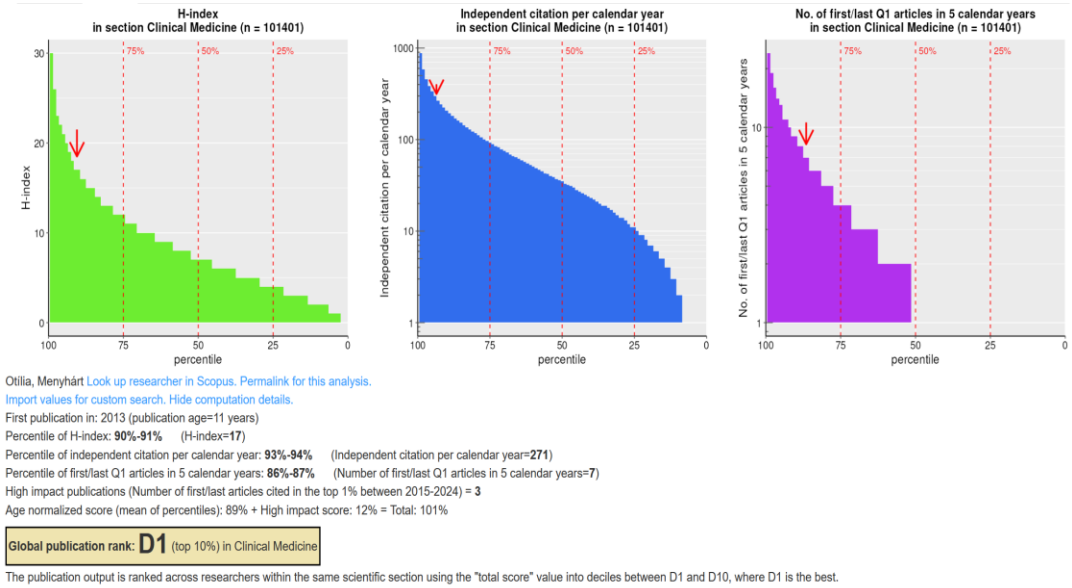


Figure 2. The ranks plot illustrates the evaluation of a randomly selected medical researcher's standing across three computed metrics: The H-index (left panel), the annual count of independent citations (central panel), and the quantity of first/last authored Q1-ranked articles in the most recent five years (right panel). Dashed red lines indicate the quartile thresholds, with red arrows pointing to the researcher's current position within these rankings.

Discussion

Evaluation of researcher output

A quantitative evaluation of researchers across disciplines is essential for uncovering hidden talents. Past performance is often a reliable indicator of future success, as demonstrated in the comparison of publication output before and after earning a PhD (Munkácsy et al. 2022). Several global platforms assess scholarly output, such as Scopus, which provides the H-index, total publications, citations, and authorship distribution; Google Scholar, which includes the H-index, total publications, and i10 index; Web of Science, which offers the H-index and citation percentiles; Semantic Scholar, which lists influential citations; and Open Alex, which reports the H-index, i10 index, and aggregate publications and citations.
However, the reliability of these metrics varies by researcher age and discipline, with younger researchers and fields like humanities seeing less accuracy.

Pilot and global application of analysis pipeline

Our pilot project, based on the Hungarian Academy of Sciences' classification system and data from the Hungarian Scientific Work Archive (HSWA) (Győrffy et al. 2022), highlighted some challenges due to the limited scope and differing categorization of HSWA data. Notably, HSWA includes shared first/last authorships and non-Scopus-indexed publications.

To address these issues, we redesigned our analysis pipeline for global applicability, using Scopus data. This redesign involved omitting corresponding authorships and adjusted filtering to include only Scopus-listed publications, avoid double-counting citations for first/last authored works, and ensure compatibility with global standards. The newly established platform now allows users to evaluate and compare the output of around 20 million researchers across various scientific fields, providing an accessible and accurate tool for global scholarly assessment.

Limitations of the approach

There are some limitations to our analysis that need to be addressed. First, Van Leeuwen et al. (2001) highlighted the language bias in citation metrics, particularly for non-English publications. Since our data source primarily includes English-language publications, researchers in non-English-speaking countries may be underrepresented. We emphasize the importance of considering these language biases when evaluating research performance at national or institutional levels.

Second, researchers often publish across multiple disciplines, which complicates bibliometric analysis. A more nuanced evaluation, incorporating contextual data, is required to accurately assess cross-disciplinary researchers. Developing methodologies to address this challenge can be a future research objective.

Third, our analysis is based on a snapshot of the database, and any changes to researchers' profiles or IDs since that snapshot—such as name similarity issues or ID cancellations due to mergers—may affect individual assessments. However, given the large number of researchers in the reference cohort, these fluctuations are unlikely to undermine the broader reliability of our approach.

Conclusions

We have introduced a new global platform for evaluating individual researchers' scientific output. Our study presents a new way of scientometric analysis, offering global coverage of 507,223 researchers across 19 disciplines and a unique methodology that accounts for publication age and disciplinary differences. The user-friendly online portal (www.scientometrics.org/scopus) democratizes complex bibliometric analysis. By integrating multiple parameters with a weighted scoring system and including a high-impact publication component, we provide a more equitable framework for assessing scientific productivity. Our method offers additional insights to complement existing evaluation practices, with our goal to ensure fair and transparent assessments within the scientific community.

References

Alchokr, R., Krüger, J., Shakeel, Y., Saake, G., & Leich, T. (2022). On academic age aspect and discovering the golden age in software engineering. Proceedings of the 15th International Conference on Cooperative and Human Aspects of Software Engineering, pp. 102–6. Presented at the ICSE 22: 44th International Conference on Software Engineering, May 21, Pittsburgh Pennsylvania: ACM.

Győrffy, B., Csuka, G., Herman, P., & Török, Á. (2020). *Is there a golden age in publication activity? — an analysis of age-related scholarly performance across all scientific disciplines, Scientometrics,* 124/2: 1081–97.

Győrffy, B., Herman, P., & Szabó, I. (2020). *Research funding: past performance is a stronger predictor of future scientific output than reviewer scores, Journal of Informetrics,* 14/3: 101050.

Győrffy, B., Nagy, A. M., Herman, P., & Török, Á. (2018). *Factors influencing the scientific performance of Momentum grant holders: an evaluation of the first 117 research groups, Scientometrics,* 117/1: 409–26.

Győrffy, B., Weltz, B., Munkácsy, G., Herman, P., & Szabó, I. (2022*). Evaluating individual scientific output normalized to publication age and academic field through the Scientometrics.org project, Methodology,* 18/4: 278–97.

Győrffy, B., Weltz, B., & Szabó, I. (2023). *Supporting grant reviewers through the scientometric ranking of applicants, PLOS ONE,* 18/1: e0280480. DOI: 10.1371/journal.pone.0280480

Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2016). *Multiple Citation Indicators and Their Composite across Scientific Disciplines, PLOS Biology*, 14/7: e1002501.

Lan, Y., Clements, K. W., & Chai, Z. K. (2023). *How Productive Are Economics and Finance PhDs? Australian Economic Review,* 56/4: 442–61.

Munkácsy, G., Herman, P., & Győrffy, B. (2022). *Comparison of scientometric achievements at PhD and scientific output ten years later for 4,790 academic researchers, PLOS ONE,* 17/7: e0271218.

Szluka, P., Csajbók, E., & Győrffy, B. (2023). *Relationship between bibliometric indicators and university ranking positions, Scientific Reports,* 13/1: 14193.

Tóth, T., Demeter, M., Csuhai, S., & Major, Z. B. (2024). *When career-boosting is on the line: Equity and inequality in grant evaluation, productivity, and the educational backgrounds of Marie Skłodowska-Curie Actions individual fellows in social sciences and humanities, Journal of Informetrics,* 18/2: 101516.

Van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Van Raan, A. F. J. (2001). *Language biases in the coverage of the Science Citation Index and its consequencesfor international comparisons of national research performance., Scientometrics*, 51/1: 335–46.

https://doi.org/10.51408/issi2025_051

Unveiling the Temporal Dynamics: The Impact of Knowledge Source Diversity, Breadth and Depth on Disruptive Innovation through Time-Series Analysis

Yue Li[1], Lele Kang[2], Jiaxing Li[3]

[1]*y.li@smail.nju.edu.cn*, [2]*lelekang@nju.edu.cn*
Laboratory of Data Intelligence and Interdisciplinary Innovation, Nanjing University, Nanjing (China)
School of Information Management, Nanjing University, Nanjing (China)

[3]*jxlee@njau.edu.cn*
School of Information Management, Nanjing Agricultural University, Nanjing (China)

Abstract

Disruptive innovation plays a critical role in driving technological progress and reshaping industries by challenging established paradigms and fostering new opportunities for growth. While previous research has largely focused on the static relationship between knowledge characteristics and disruptive innovation, the temporal evolution of knowledge source diversity, breadth, and depth and their influence on disruptive innovation remain unclear. This study explores these dynamics by analysing multivariate time-series data from global patents spanning 1980 to 2010. The Autoregressive Distributed Lag (ARDL) model is employed to assess both the short-run and long-run effects of knowledge structures on disruptive innovation. The results reveal that, in the long run, knowledge source diversity positively influences disruptive innovation, whereas knowledge breadth has a negative effect, and knowledge depth shows no significant impact. In the short run, knowledge depth positively contributes to innovation, while knowledge source diversity exerts a negative effect, and knowledge breadth remains insignificant. These findings underscore the importance of aligning knowledge management strategies with temporal dynamics to foster sustained innovation.

Introduction

Disruptive innovation, which reshapes existing technological paradigms and drives progress in entirely new directions, has historically been a cornerstone of transformative development. However, recent studies reveal a worrying trend: the disruptive potential of innovations is steadily declining. Park et al. (2023) quantified this phenomenon using the CD index, a metric that captures the disruptiveness of patents and scientific publications by assessing their impact on subsequent citation patterns. Their findings highlighted a consistent decline in disruptiveness across technological fields, raising critical questions about the factors driving this shift. Despite growing attention to this phenomenon, it remains unclear whether and how different dimensions of knowledge influence this decline.

Existing studies have investigated various factors influencing innovation, including institutional frameworks such as intellectual property rights regime (Thakur-Wernzet al., 2022) and funding mechanisms (Irfan et al., 2022), technological ecosystems such as industry clusters (Kim et al., 2023) and R&D networks (Wen et al., 2021), and organizational characteristics such as team size (Wuchty et al., 2007), leadership styles (Alblooshi et al., 2021), and knowledge management practices (Darroch, 2005; Mardani et al., 2018). Among these factors, knowledge emerges as a cornerstone of

the innovation process, enabling both exploration and exploitation, which form the basis for novel recombination and technical refinement (Grant, 1996). Evolutionary economics reinforces this perspective by emphasizing the cumulative nature of knowledge, where its recombination drives breakthroughs (Nelson, 1985). Despite these insights, most research adopts a static perspective, overlooking how the continuous evolution of knowledge influences disruptive innovation. Innovation is inherently dynamic, shaped by the transformation of knowledge and its interplay with external factors like technological advancements and market dynamics. As innovation systems mature, the complexity of integrating and applying knowledge evolves, potentially reshaping its impact on innovation outcomes. This highlights the need to examine how the dynamic restructuring of knowledge affects the trajectory of disruptive innovation.

Innovation does not occur in isolation; it is inherently shaped by the knowledge that drives it (Kaplan et al., 2015). From the perspective of the knowledge-based view, the evolution of disruptive innovation is fundamentally shaped by two critical dimensions of knowledge: *what knowledge is combined*, referring to Knowledge Source Diversity (KSD), and *how knowledge is applied*, referring to Knowledge Breadth (KB) and Depth (KD) (Grant, 1996). *what knowledge is combined* pertains to the sources of knowledge that contribute to an innovation, capturing the diversity of external knowledge inputs that provide the raw material for technological advancement. In contrast, *how knowledge is applied* focuses on the internal structuring and utilization of knowledge within the innovation process, reflecting the breadth and depth with which knowledge is synthesized and leveraged to achieve disruptive breakthroughs. Specifically, KSD refers to the variety of origins from which knowledge is drawn, including different technological domains, industries, and institutional sources. A high degree of KSD fosters novel recombination and cross-boundary integration, introducing fresh perspectives that challenge established paradigms (Rodriguez et al., 2017). However, the complexity of assimilating and coordinating diverse external knowledge inputs can impose integration challenges, potentially delaying the realization of innovation benefits. KB and KD, representing the *how* dimension, determine how acquired knowledge is internally structured and applied within an innovation. KB reflects the degree of interdisciplinarity within a single innovation effort. Greater KB facilitates the integration of diverse ideas, fostering interdisciplinary breakthroughs; however, it can also introduce internal coordination complexities that may hinder short-run efficiency. In contrast, KD signifies the extent of specialization within a particular domain, enabling focused technical advancements that build upon existing expertise. While deep specialization supports incremental innovation and enhances technical proficiency, it may limit adaptability and reduce the potential for radical disruption over time.

To better understand the dynamic relationship between disruptive innovation (DI) and the three critical dimensions of knowledge—source diversity, breadth, and depth—this study employs the Autoregressive Distributed Lag (ARDL) model (Pesaran, et al., 2001). In contrast to traditional static models, the ARDL approach enables the simultaneous estimation of short-run adjustments and long-run equilibrium relationships, providing deeper insights into the evolving impact of knowledge on DI. By distinguishing between short-run fluctuations and long-run

trends, the ARDL model offers valuable insights into how DI responds to changes in knowledge dimensions over different time horizons. The short-run analysis reveals immediate responses to shifts in knowledge, while the long-run analysis captures persistent influences that shape innovation trajectories. This comprehensive approach contributes to a deeper understanding of how knowledge recombination and application influence DI.

In order to validate our findings, this study analyses annual patent data from 1980 to 2010. Unit root tests, including the Augmented Dickey-Fuller and Phillips-Perron tests, are applied to ensure the stationarity of the variables. Given the mixed integration order commonly found in time-series data, the ARDL bounds test is applied to determine the presence of long-run relationships between DI and the knowledge examined in this study. The findings indicate that, over the long run, a higher diversity of knowledge sources enhances disruptive innovation, whereas broader knowledge integration has an adverse effect, and the influence of knowledge depth is not statistically significant. In the short run, increased knowledge depth plays a positive role in fostering disruptive innovation, while greater knowledge source diversity presents challenges, and knowledge breadth does not exhibit a noticeable impact.

Building on these findings, this study makes several contributions to the literature. First, they provide a deeper understanding of the mechanisms underlying the observed decline in disruptiveness, highlighting the lack of sufficient analysis on the temporal evolution of knowledge structures. Second, by employing the ARDL model, this study offers a methodological advancement that allows for the investigation of DI from a dynamic perspective, capturing both short-run adjustments and long-run equilibrium relationships. Third, the study provides actionable insights for policymakers and innovation managers by emphasizing the importance of balancing knowledge diversity, breadth, and depth across different time horizons to foster sustained disruptive innovation.

Related Work

The theory of disruptive innovation was first proposed by Christensen (1997), characterized by its non-linear technological trajectory. Unlike traditional mainstream technologies, disruptive innovation advances through differentiated strategies to achieve competitive advantage (Hang et al., 2015). Existing studies have defined the concept from various perspectives, including technological characteristics (Nagy et al., 2016; Reinhardt and Gurtner, 2015), innovation processes (Levina, 2017), and innovation impacts (Suseno, 2018). These studies have also explored disruptive innovation across multiple levels, including the individual (Osiyevskyy and Dewald, 2015), firm (Van Balen et al., 2019), industry (Chevalier-Roignant et al., 2019), and network/ecosystem levels (Ruan et al., 2014). Despite widespread attention from academia and practice, the core concept of disruptive innovation remains ambiguous and inconsistent, which limits the development of the theory. Specifically, the mechanisms of disruptive innovation in technological contexts and its relationship with knowledge structures require further exploration.

Knowledge structure, as a critical driver of innovation, is commonly described through two dimensions: knowledge breadth and depth. These dimensions constitute

key elements of the knowledge base. Knowledge breadth refers to the extent to which a patent integrates knowledge from multiple fields, reflecting the degree to which diverse ideas are synthesized within the innovation itself. In contrast, knowledge depth represents specialized expertise within a specific field, emphasizing the sophistication of technological development (Zou et al., 2019). Existing research indicates that knowledge breadth facilitates innovation, particularly disruptive innovation, by enabling diverse combinations of technologies and cross-domain integration (Xu et al., 2015). However, excessive knowledge breadth may lead to resource dispersion and coordination complexities, thereby hindering innovation efficiency (Jin et al., 2015). In contrast, knowledge depth strengthens technological advantages in specific fields, supporting incremental innovation (Boh et al., 2014). Yet, over-reliance on knowledge depth may limit adaptability to emerging technologies, particularly in rapidly changing technological environments.

Knowledge source diversity introduces an external driving force for technological innovation. On one hand, diverse knowledge sources enrich opportunities for technological combinations and enhance innovation capacity. For instance, Dogru et al. (2019) highlighted that integrating knowledge from different sources significantly improves innovation performance, especially in resource-constrained contexts. Additionally, knowledge source diversity provides the necessary resilience and adaptability for disruptive innovation, enabling technical systems to address path dependency and uncertainties (Luo et al., 2024). On the other hand, excessive diversity in knowledge sources may increase coordination challenges and integration costs, thereby negatively impacting innovation efficiency. Hajialibeigi (2023) identified an inverted U-shaped relationship between knowledge source diversity and innovation performance, where moderate diversity optimizes resource utilization while excessive diversity exacerbates management complexity. Furthermore, the impact of knowledge source categories on technological innovation differs significantly. Abdul Basit and Medase (2019) demonstrated that public sector knowledge better promotes technological innovation in manufacturing, whereas private sector knowledge integration is more effective in service industries.

Data and Method

Data and variables

To investigate the short-run and long-run dynamics between disruptive innovation, knowledge source diversity, breadth and depth, this study utilizes patent data obtained from the PatentView database. This comprehensive database includes detailed information on patents from 1976 to 2024, encompassing inventor details, patent and application metadata, assignee and location information, as well as International Patent Classification (IPC) data.

The database further provides access to the full text of patents, which includes three key sections: abstract, claims, and description. The claims section outlines the scope of the legal protection granted to the patent, while the description section provides a detailed explanation of the invention or innovation’s technical characteristics. The abstract offers a summary of the content in both the claims and description sections.

To analyse the genuine technological attributes of patented inventions, this study exclusively relies on the description section.

Disruptive Innovation. Disruptive innovation is measured using the CD index, which was developed by Funk and Owen-Smith (2017) and later applied by Park et al. (2023). The CD index quantitatively captures whether a patent consolidates existing knowledge or disrupts the technological status quo. Consolidating patents build upon prior knowledge and reinforce established trajectories, whereas disruptive patents render earlier work obsolete and chart new technological directions. The CD index ranges from -1 to 1, where -1 indicates a highly consolidating innovation, and 1 signifies a highly disruptive innovation.

This study adopts the five-year post-publication window used by Park et al. (2023), referred to as CD_5, to evaluate the disruptive potential of patents. The starting year of analysis is 1980, aligns with Park et al.'s dataset to ensure consistency in the time window and methodology. The calculation of the CD index also follows the exact formula proposed by Park et al. (2023). Using this standardized approach ensures comparability with prior studies and allows for robust exploration of the relationships between disruptive innovation and knowledge dimensions, including breadth, depth, and source diversity.

Knowledge Source Diversity. The Knowledge Source Diversity (KSD) measures the extent to which a patent integrates knowledge from multiple technological categories, based on the NBER two-digit technology classification. The NBER classification system, developed by Hall et al. (2001), provides a standardized framework for categorizing patents into broad technological fields, facilitating cross-field comparisons, and enabling robust analyses of knowledge diversity. In this study, the classification of patents into NBER technology categories is obtained directly from the PatentView database, ensuring consistency and reliability in the analysis. To calculate KSD, the references cited by each patent are analysed to determine their distribution across NBER technology categories. The diversity of these references is quantified using an entropy-based approach, which accounts for both the number of categories referenced and the balance among them. Patents with higher KSD indicate a greater reliance on knowledge inputs from multiple distinct technological fields, reflecting their ability to integrate diverse sources of knowledge. This diversity is hypothesized to enhance the potential for creative recombination of ideas, which is often a critical driver of disruptive innovation.

Knowledge Breadth. The Knowledge breadth (KB) is defined as the extent to which a patent draws upon vocabulary from multiple technological fields. Following the methodology outlined in Bowen et al. (2023), this metric is constructed by first calculating the frequency of word usage across technological fields for each year. A word is tagged as *specialized* in a particular field if its usage in that field exceeds 150% of its usage in the second most prominent field during the same year. Words that do not meet this criterion are classified as *unspecialized* and excluded from further analysis. For each patent, the fraction of specialized words classified into each field is then calculated, with these fractions summing to one for every patent. Using this classification, technological breadth is defined as one minus the concentration of specialized words, thereby reflecting the diversity of fields from which a patent draws

its vocabulary. Patents with high knowledge breadth integrate terminology from a wider range of fields, indicating a more diverse knowledge base:

Knowledge Depth. The Knowledge Depth (KD) measures the extent of focus within a single technological field, and is calculated based on the concentration of a patent's classification within a specific four-digit International Patent Classification (IPC4) code. The IPC4 system provides a highly granular framework for categorizing patents, often used as a proxy for defining technological fields. By examining the proportion of a patent's classifications that fall within its most dominant IPC4 category, knowledge depth captures the degree to which a patent concentrates on a single technological field. Patents with high knowledge depth often exhibit a deliberate emphasis on advancing a particular field, suggesting a refined specialization that may impact incremental innovations or significant technical improvements within that domain. By anchoring the measurement of depth in the IPC4 classification, the analysis ensures precision in capturing the technical focus of each patent. This reliance on established knowledge structures may enhance efficiency in knowledge utilization. All variables and their description are shown in Table 1.

Table 1. Variables description.

Variables	*Description*
CD	Measured using the CD index, developed by Funk and Owen-Smith (2017) and applied by Park et al. (2023). The index ranges from -1 (highly consolidating) to 1 (highly disruptive), with CD_5 calculated over a five-year post-publication window to evaluate a patent's influence on obsolescing or reinforcing prior knowledge.
Knowledge Source Diversity (KSD)	Reflects the variety of technological categories from which a patent integrates knowledge. Based on the NBER two-digit technology classification and calculated using entropy to measure the diversity of references cited by each patent across multiple fields.
Knowledge Breadth (KB)	Captures the diversity of technological fields from which a patent draws its vocabulary. Calculated as one minus the concentration of a patent's classification into six broad fields, reflecting the extent to which the patent spans multiple domains. Derived using field-specific data from the patent text and classification systems.
Knowledge Depth (KD)	Measures the extent of focus within a single technological field. Calculated based on the proportion of a patent's classifications concentrated within its most dominant IPC4 code, representing a refined specialization in a specific domain.

The rationale for employing distinct operational measures for KSD, KB, and KD is grounded in their theoretical separation, empirical complementarity, and granular alignment with the conceptual constructs. Although these dimensions are interrelated, they reflect fundamentally different structural layers of knowledge, which necessitates differentiated yet coherent measurement strategies. First, KSD captures the diversity of technological origins, for which the NBER 2-digit classification is particularly suited. Its coarse granularity reflects broader source fields (e.g., Chemicals**,** Electronics, Drugs), and has been widely used to proxy knowledge origin

variety in macro-level innovation studies (Hall et al., 2001). NBER codes aggregate IPC-based patent classes according to economically meaningful technological sectors, thus aligning closely with the idea of where knowledge comes from. Second, KD is intended to reflect technological specialization, which demands greater classification precision. The IPC 4-digit level provides such fine-grained technical delineation, enabling us to observe how concentrated a patent's technical focus is. Compared with higher-level IPC or NBER codes, IPC4 provides domain stability and domain resolution, making it the most valid proxy for focused depth within a technological field. Third, KB concerns the semantic recombination and interdisciplinary expression of knowledge within the patent text. To this end, a vocabulary-based approach is employed, tracking the field-specific concentration of technical terms used in abstracts and claims. This textual metric captures horizontal conceptual integration at a finer level than taxonomic classifications, especially in domains where innovation involves hybrid or emergent concepts not yet classified in IPC/NBER systems. While the data sources and granularity differ across these three variables, they are intentionally selected to match the theoretical domain of each construct: broad origin domains (KSD), fine technical depth (KD), and semantic conceptual spread (KB). These differences do not imply inconsistency but rather reflect the layered nature of knowledge structures in innovation. We explicitly acknowledge that the classification schemes are non-nested and differ in dimensional logic. However, their temporal aggregation into annual panel data and their independent derivation from non-overlapping sources reduce concerns about collinearity or semantic redundancy. Moreover, our ARDL model framework allows for distinct lag structures, further reducing risks of artificial convergence.

Methodology and model specification

Econometric methods that investigate the temporal dynamics of innovation processes are essential for understanding how variables interact over time. These approaches enable the analysis of both short-run fluctuations and long-run equilibrium relationships, offering valuable insights into the mechanisms impacting disruptive innovation and its connections to knowledge dimensions such as source diversity, breadth and depth. Given the need to examine these dynamics comprehensively, this study adopts the Autoregressive Distributed Lag (ARDL) bounds testing model, introduced by Pesaran et al. (1999) and later developed further (Pesaran, et al., 2001), to explore the cointegration processes and temporal interactions among the variables. The ARDL approach not only estimates cointegration and long-run equilibrium relationships but also captures dynamic effects in both time horizons, offering a comprehensive framework for understanding temporal interactions.

The ARDL methodology is particularly advantageous for several reasons. First, it is highly flexible and can accommodate variables with mixed integration orders, whether I (0) or I (1). Second, the single-equation setup simplifies implementation and interpretation compared to traditional cointegration methods. Third, it allows for different lag lengths to be specified for different variables, enhancing the model's adaptability to the data. Fourth, the method is well-suited for small sample sizes, providing robust estimates of long-run relationships and parameters. Finally, the ARDL model effectively addresses potential issues of autocorrelation and

endogeneity, ensuring unbiased and reliable results (Harris and Sollis, 2003; Jalil and Ma, 2008).

Given these strengths, the ARDL approach is employed in this study to examine the temporal dynamics between disruptive innovation and its key regressors, such as knowledge source diversity, breadth and depth. The method is applied to identify both the long-run equilibrium relationships and the short-run adjustments that occur in response to deviations from equilibrium. The subsequent steps for verifying these dynamics within the ARDL framework are outlined in the following sections.

Stationarity test. Stationarity is a critical consideration in time-series analysis, as it ensures the validity of econometric models and the reliability of their results. Time-series data have diverse applications across various fields, and identifying the appropriate trend structure of the data represents an essential econometric task (Mushtaq, 2011). To determine the stationarity of the variables, this study employs the Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) unit root tests. These tests are widely used to identify whether variables are stationary at their levels or become stationary after differencing. The results of these tests guide the appropriate application of the Autoregressive Distributed Lag (ARDL) approach, which is capable of handling variables integrated at different orders. Specifically, the ARDL model can accommodate variables that are stationary at level (I(0)), at first difference (I(1)), or a combination of the two, making it a robust method for analysing the cointegration and temporal dynamics among time-series variables.

Autoregressive Distributed Lag bounds test. The bounds testing procedure is utilized in this study to examine whether a single long-run relationship exists among the variables under investigation. The ARDL bounds test evaluates cointegration by testing the joint significance of the coefficients of the lagged levels of the variables in a single-equation model. The model for the bounds test is specified as follows:

$$\Delta CD_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta CD_{t-i} + \sum_{i=0}^{q} \gamma_i \Delta KB_{t-i} + \sum_{i=0}^{r} \delta_i \Delta KD_{t-i} + \sum_{i=0}^{s} \eta_i \Delta KSD_{t-i} + \theta_1 CD_{t-1} + \theta_2 lnKB_{t-1} + \theta_3 lnKD_{t-1} + \theta_4 lnKSD_{t-1} + \epsilon_t$$

In this equation, Δ denotes the first-difference operator, CD_t is the disruptive innovation index, and KB_t, KD_t, and KSD_t represent knowledge breadth, depth, and source diversity, respectively. The optimal lag lengths (p, q, r, s) are determined using the Akaike Information Criterion (AIC), which minimizes information loss and ensures the model is parsimonious while retaining explanatory power. The coefficients θ_1, θ_2, θ_3, θ_4 capture the long-run equilibrium relationships, while the summations account for short-run dynamics. The term ϵ_t captures any variations unexplained by the model, ensuring the robustness of the estimation process.

To evaluate the existence of a cointegration relationship, the ARDL bounds test is applied. This test compares the calculated F-statistic to critical bounds for the null hypothesis (H_0), which assumes no cointegration among the variables, and the alternative hypothesis (H_1), which posits the presence of cointegration. A rejection of H_0 occurs when the F-statistic exceeds the upper critical bound, indicating a stable long-run relationship among the variables. Conversely, if the F-statistic falls below

the lower bound, the null hypothesis cannot be rejected. When the F-statistic lies between the bounds, the result is inconclusive, requiring further investigation.

Once a long-run relationship is confirmed through the ARDL bounds testing approach, the model is re-specified into an Error Correction Model (ECM) to estimate both short-run dynamics and the speed of adjustment back to the long-run equilibrium. The ECM effectively integrates short-run fluctuations and long-run relationships within a single framework, ensuring the model captures both immediate and equilibrium effects of the independent variables on disruptive innovation. The ECM for this study is specified as follows:

$$\Delta CD_t = \alpha + \sum_{i=1}^{p} \beta_i \Delta CD_{t-i} + \sum_{i=0}^{q} \gamma_i \Delta KB_{t-i} + \sum_{i=0}^{r} \delta_i \Delta KD_{t-i} + \sum_{i=0}^{s} \eta_i \Delta KSD_{t-i} + \tau ECT_{t-1} + \epsilon_t$$

The ECM framework is particularly valuable because it allows the separation of short-run dynamics from long-run equilibrium behaviour while maintaining a consistent representation of the temporal relationships among variables. The short-run effects are captured by the coefficients of the lagged differences, which provide insights into the immediate impacts of changes in knowledge dimensions on disruptive innovation. Meanwhile, the Error Correction Term (ECT) integrates the short-run adjustments with the long-run relationship, ensuring that deviations from equilibrium are systematically corrected over time.

By applying the ECM within the ARDL framework, this study is able to investigate not only how knowledge breadth, depth, and source diversity influence disruptive innovation in the long run, but also how these variables interact dynamically in the short run. This dual focus provides a comprehensive understanding of the temporal mechanisms impacting innovation processes.

Stability test. Ensuring the stability of regression models is critical when working with autoregressive structures, as stability confirms the robustness of estimated coefficients over time. In this study, the CUSUM of squares approach, as proposed by Brown et al. (1975), is employed to evaluate the dynamic stability of the model. The CUSUM of squares test provides a graphical representation of stability, where the plotted test statistic is compared against a confidence interval. If the test statistic remains within the confidence bounds, the model is considered stable, indicating no significant changes in the regression coefficients over time. Conversely, if the statistic crosses the bounds, it suggests potential instability, requiring further investigation.

Empirical findings

This study employs multivariate time-series data from 1980 to 2010, with annual observations to mitigate the influence of seasonal variations. The annual data are derived by calculating patent-level indicators for each year and then averaging these values at the yearly level, ensuring a consistent representation of trends over time. The analysis focuses on identifying the relationships between disruptive innovation and various knowledge dimensions over time.

Summary statistics

The descriptive statistics for the key study variables is provided in Table 2, including disruptive innovation (CD), knowledge breadth (lnKB), knowledge depth (lnKD), and knowledge source diversity (lnKSD). The mean value of CD is 0.127, with a standard deviation of 0.098, indicating moderate variation in disruptive innovation across the sample period. The minimum and maximum values of CD range from 0.030 to 0.388, reflecting substantial differences in the disruptiveness of innovations over time. Knowledge breadth (lnKB) exhibits a mean value of 0.364 with relatively low variability (S.D. = 0.023), suggesting a consistent level of knowledge integration across patents. Knowledge depth (lnKD) has a slightly higher mean of 0.528 and also demonstrates low variability (S.D. = 0.017), highlighting the stable specialization within individual technological fields. In contrast, knowledge source diversity (lnKSD) shows the highest mean of 0.684 with minimal variation (S.D. = 0.003), indicating that patents consistently rely on a diverse set of external knowledge sources.

Table 2. Summary statistics of study variables.

Variables	*Mean*	*S.D.*	*Min*	*Max*
CD	0.127	0.098	0.030	0.388
lnKB	0.364	0.023	0.313	0.397
lnKD	0.528	0.017	0.488	0.561
lnKSD	0.684	0.003	0.677	0.687

Together, these descriptive statistics and time trends highlight the dynamic relationships between disruptive innovation and the key knowledge dimensions, providing a foundation for exploring their short-run and long-run interactions in subsequent analyses.

Stationarity test

In this study, stationarity of the variables was tested using both the Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) tests, with the results summarized in Table 3. The stationarity test results reveal that, except for the variable CD, all variables become stationary after applying the first difference. Specifically, the results indicate that at the level, none of the variables, except for CD, exhibit stationarity. However, after taking the first difference, all variables—namely the logarithms of knowledge breadth (lnKB), knowledge depth (lnKD), and knowledge source diversity (lnKSD)—become stationary. The variable CD, on the other hand, is stationary at the level, confirming that it does not require differencing. This mixed order of integration among the variables suggests that an Autoregressive Distributed Lag (ARDL) bound approach is appropriate for modeling the relationship between

the variables, as it can accommodate variables with different integration orders (i.e., I (0) and I (1)).

Table 3. Stationarity test statistics.

Variables	*ADF Test*		*PP Test*		*Stationary Remark*
	Level	*First difference*	*Level*	*First difference*	
CD	-4.873*** (0.000)	-	-3.870*** (0.013)	-	I (0)
lnKB	-0.191 (0.992)	-5.894*** (0.000)	-0.196 (0.992)	-5.872*** (0.000)	I (1)
lnKD	-2.676 (0.246)	-4.637*** (0.001)	-2.722 (0.227)	-4.574*** (0.001)	I (1)
lnKSD	-1.596 (0.794)	-7.120*** (0.000)	-1.325 (0.882)	-7.105 (0.000)	I (1)

Note: An intercept term and a trend term have been included in all unit-root tests. Significance levels are denoted as 1%, 5%, and 10% with ***, **, and * respectively.

ARDL bounds test

To determine the optimal lag length for the model, the Akaike Information Criterion (AIC) was utilized. Based on this criterion, the chosen model is ARDL (1, 0, 2, 2). This means that the optimum lag lengths for the variables CD, lnKB, lnKD, and lnKSD are p=1, q=0, r=2 and s=2, respectively. The results of the ARDL bounds test are presented in Table 4, which includes the F-statistics values for testing the presence of a long-run relationship between the variables.

Table 4. ARDL bounds test (F-statistic).

F-statistic		*Null hypothesis: no levels of relationship*		
	Value	*Significance level*	*I (0)*	*I (1)*
Value of F-statistic	31.232	10.0%	2.72	3.77
K	3	5.0%	3.23	4.35
Critical Value Bounds	0.1-0.01	2.5%	3.69	4.89
		1.0%	4.29	5.61

Since the F-statistic value exceeds the critical values for both I (0) and I (1), this provides strong evidence of a long-run relationship among the variables. The results suggest that the knowledge dimensions (KB, KD, KSD) are jointly influencing

disruptive innovation in the long-run, while the variables move together toward an equilibrium over time.

ARDL adjustment estimation, long-run and short-run relationships

The ARDL adjustment estimates is reported in Table 5, indicating how the variables align with the long-run equilibrium following deviations. The coefficient of CD L1 is −0.135, which is negative and statistically significant at the 1% level. This value reflects the proportion of the adjustment toward long-run equilibrium in response to deviations. Specifically, approximately 13.5% of the disequilibrium is corrected within one year, indicating that the variables are gradually realigned with their long-run equilibrium. The statistically significant negative coefficient also suggests a stable long-run relationship, with adjustments occurring systematically over time.

Table 5. ARDL adjustment estimates.

D.CD	*Coef.*	*Std.error*	*T*	*P >|t|*	*[95% Conf. Interval]*	
CD. L1.	-0.135***	0.023	-5.98	0.00	-0.181	-0.088

Note: Significance levels are denoted as 1%, 5%, and 10% with ***, **, and * respectively.

The long-run estimates obtained from the ARDL model is presented in Table 6, illustrating the sustained relationships between disruptive innovation and the knowledge dimensions: breadth, depth, and source diversity. The coefficient of knowledge breadth (lnKB) is negative and statistically significant at the 1% level. This indicates that in the long run, an increase in knowledge breadth is associated with a reduction in disruptive innovation. This may reflect the trade-off between generalization and specialization, where increased knowledge breadth could dilute the focus needed for achieving disruptive breakthroughs. The coefficient of knowledge depth (lnKD) is negative but not statistically significant. This result implies that knowledge depth does not show a strong long-run influence on disruptive innovation during the study period. This finding may suggest that depth alone is insufficient to drive innovation without the complementary effects of breadth or diversity. The coefficient of knowledge source diversity (lnKSD) is positive and statistically significant at the 1% level. This indicates a strong positive long-run relationship between knowledge source diversity and disruptive innovation. The result suggests that integrating diverse sources of knowledge significantly enhances the potential for disruptive breakthroughs, potentially due to the cross-pollination of ideas from different fields or disciplines.

Table 6. ARDL long-run estimates.

Variables	*Coef.*	*Std.error*	*T*	*P >|t|*	*[95% Conf.*	*Interval]*
lnKB	-1.276***	0.396	-3.23	0.004	-2.101	-0.451
lnKD	-0.558	0.988	-0.57	0.578	-2.619	1.503
lnKSD	18.942***	4.209	4.50	0.000	10.161	27.722

Note: Significance levels are denoted as 1%, 5%, and 10% with ***, **, and * respectively.

Table 7 reports the short-run estimates from the ARDL model, capturing the immediate effects of knowledge dimensions on disruptive innovation. The results indicate that the variable knowledge breadth (lnKB) does not return significant short-run coefficients, suggesting that it may not play a measurable role in influencing disruptive innovation within the short-run time horizon. This lack of significant results could be attributed to the inherently gradual nature of the effects of knowledge breadth, which may require longer periods to manifest its impact on innovation outcomes. For knowledge depth (lnKD), the results reveal a positive and statistically significant short-run relationship with disruptive innovation. At lag order 0, the coefficient is 0.270, significant at the 1% level, indicating that an immediate increase in knowledge depth is associated with a rise in disruptive innovation. This positive relationship persists at lag order 1, with a smaller coefficient of 0.185, which is significant at the 10% level. These findings suggest that while knowledge depth contributes positively to disruptive innovation in the short run, the magnitude of its impact diminishes slightly over time. In contrast, knowledge source diversity (lnKSD) shows a consistently negative and statistically significant short-run relationship with disruptive innovation. At lag order 0, the coefficient is -4.829, significant at the 1% level, indicating that an increase in knowledge source diversity imposes short-run challenges on innovation processes. This negative impact persists at lag order 1, with a coefficient of -4.953, also significant at the 1% level. The consistent short-run negative effects of knowledge diversity suggest that the integration of diverse knowledge sources may introduce complexities and inefficiencies that hinder immediate innovation outcomes, despite its positive influence in the long run. The overall model demonstrates a strong fit, as reflected by the R-squared value of 0.936, which indicates that 93.6% of the variation in disruptive innovation can be explained by the short-run dynamics of the model.

Table 7. ARDL short-run estimates.

Variables	*Coefficient*	*Estimates*
Lag order	0	1
ΔlnKB	-	-
ΔlnKD	0.270*** (0.011)	0.185* (0.092)
ΔlnKSD	-4.829*** (0.001)	-4.953*** (0.000)
R^2	0.936	

Note: Short-run estimators for first lagged have been depicted by Δ. Significance levels are denoted as 1%, 5%, and 10% with ***, **, and * respectively.

Stability test findings

The cumulative sum of squares (CUSUM square) plot is illustrated in Figure 1, which is used to assess the stability of the regression coefficients in the specified model. The test was conducted with a 5% significance level, and the shaded area represents the confidence interval under the null hypothesis of stability. The red plot line indicates the recursive cumulative sum of squares. The stability of the model is determined by examining whether the red plot line remains within the shaded confidence bands throughout the observation period. As shown in Figure 1, the cumulative sum of squares stays entirely within the 95% confidence interval. This confirms that there is no significant deviation from stability over the study period. At the 5% significance level, the results provide evidence of the stability of the regression coefficients. The findings indicate that the model is robust and the relationships among the variables remain consistent over time.

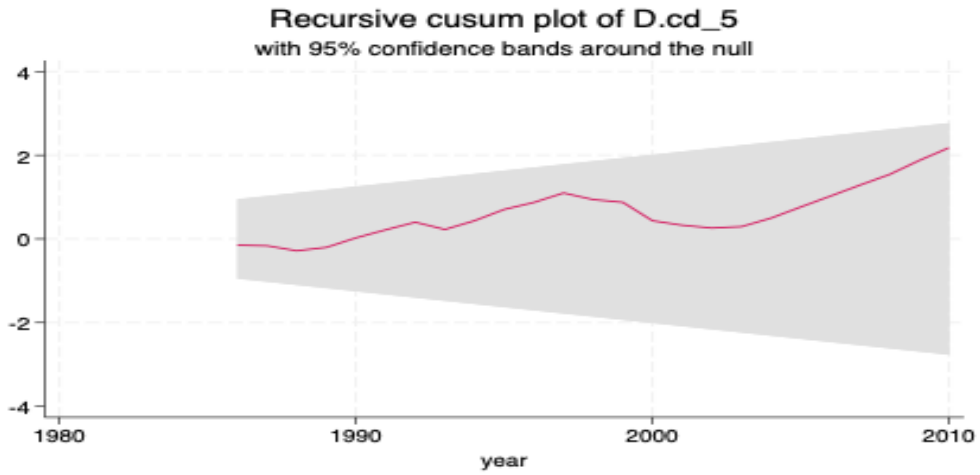


Figure 1. CUSUM Squares Plot with a 5 % level of significance.

Discussion

The findings of *what knowledge is combined* (knowledge source diversity) and *how knowledge is applied* (knowledge breadth and depth) reveal dynamic and time-dependent patterns in their effects on innovation. Knowledge source diversity, representing the richness of external inputs, negatively impacts innovation in the short run, reflecting integration challenges, yet demonstrates significant positive effects in the long run, highlighting its transformative potential. In contrast, knowledge breadth and depth, which capture the internal application of knowledge, present opposite dynamics: breadth remains insignificant in the short term but negatively influences innovation over time, while depth fosters short-run advancements but loses its significance in the long run. These seemingly paradoxical results raise important questions about the temporal trade-offs and interactions between external diversity and internal application, providing the foundation for a deeper analysis of the mechanisms underlying these patterns.

Table 8. Long-run and short-run effects of different variables.

Dependent Variable: CD	*Long-run estimate*	*Short-run estimate*
lnKB	Significant negative	-
lnKD	-	Significant positive
lnKSD	Significant positive	Significant negative

Focused paths or fragmented horizons: the temporal trade-offs of leveraging knowledge

The contrasting short- and long-run effects of knowledge breadth and depth reveal the dynamic complexities of how knowledge is leveraged in impacting innovation. In the short run, knowledge depth emerges as a significant positive factor, underscoring the power of specialization to provide focused pathways for immediate technical advancements. By concentrating resources within specific fields, depth enables the swift resolution of technical challenges and accelerates innovation within well-defined domains. However, over time, this very focus can lead to diminishing returns, as excessive specialization restricts adaptability and reduces opportunities for cross-domain exploration, ultimately limiting its long-run influence on innovation.

Conversely, knowledge breadth shows no significant impact in the short run, suggesting that the integration of diverse knowledge inputs often requires time to coordinate. Yet, in the long run, breadth exhibits a negative effect, pointing to the potential pitfalls of excessive diversification. While broader knowledge integration holds promise for fostering interdisciplinary breakthroughs, it also increases the complexity of coordination and the risk of resource fragmentation. Over time, these challenges may outweigh the benefits, resulting in innovations that are incremental rather than disruptive. This temporal trade-off highlights the critical balance required between specialization and diversification to optimize innovation outcomes over different time horizons.

A double-edged sword: the temporal dynamics of knowledge source diversity

The dual impacts of knowledge source diversity (KSD) on innovation over the short and long run highlight its role as both a catalyst and a challenge. In the short term, KSD exhibits a significant negative effect, suggesting that the inherent complexity of integrating diverse external knowledge sources can temporarily hinder innovation. This may arise from the increased coordination costs, alignment challenges, and the need for firms or inventors to navigate conflicting perspectives and methodologies. Such complexities often delay the realization of tangible innovation benefits, creating a temporal "integration burden" that suppresses short-run performance.

In contrast, the long-run positive impact of KSD underscores its transformative potential once integration barriers are overcome. Diverse knowledge sources enrich the innovation process by introducing novel ideas, fostering cross-boundary synergies, and enabling adaptability to changing technological and market landscapes. Over time, these benefits accumulate, impacting breakthroughs that are less likely to emerge from homogenous or narrowly focused knowledge pools. This positive effect reflects the delayed yet powerful rewards of leveraging external diversity, as innovation systems adapt to complexity and transform it into a source of competitive advantage.

The contrasting short- and long-run effects of KSD illustrate the importance of temporal dynamics in understanding the innovation process. While diversity can impose short-run costs, its long-run benefits reveal the necessity of investing in mechanisms that facilitate the effective integration and utilization of heterogeneous

knowledge sources. This *double-edged sword* demands strategic foresight to balance the immediate challenges with the long-run opportunities it affords.

Internal Breadth vs. External Diversity: divergent long-run paths to innovation

The contrasting long-run effects of knowledge breadth (KB) and knowledge source diversity (KSD) underscore their fundamentally different mechanisms in shaping innovation outcomes. While both dimensions represent forms of diversity, their influence diverges due to the distinct ways they interact with innovation systems over time.

Knowledge breadth, rooted in the internal integration of diverse knowledge fields within a patent, exerts a negative long-run impact on innovation. This outcome suggests that an overly broad internal knowledge base can lead to resource dispersion and coordination challenges that dilute focus. As the complexity of managing disparate knowledge fields grows, the innovation process may become fragmented, resulting in incremental improvements rather than disruptive breakthroughs. The negative effect of KB highlights the inherent difficulty of maintaining coherence and depth when attempting to integrate too many diverse internal elements over extended periods.

In contrast, knowledge source diversity, which reflects the richness of external inputs, exhibits a significant positive impact in the long run. This result points to the cumulative advantages of drawing from diverse external knowledge sources, which enrich the innovation process by introducing novel perspectives and fostering cross-boundary synergies. Unlike internal breadth, external diversity benefits from the broader ecosystem's adaptability and collaborative potential. Over time, organizations and inventors are better able to overcome the initial challenges of integrating diverse sources, transforming external complexity into a platform for sustained innovation and adaptability to emerging trends.

The divergent long-run effects of KB and KSD highlight the critical distinction between internal and external diversity. While internal breadth often struggles with the constraints of resource allocation and focus, external diversity thrives on the dynamism of collaborative ecosystems and the ability to recombine knowledge from varied origins. Understanding these differences underscores the importance of aligning knowledge strategies with the unique demands of long-run innovation, leveraging external diversity to complement and counterbalance the limitations of internal breadth.

Conclusion

In recent decades, the innovation landscape has undergone profound changes driven by increasingly complex knowledge structures. This study contributes to a more dynamic understanding of how knowledge source diversity (KSD), breadth (KB), and depth (KD) influence disruptive innovation over time. By applying an Autoregressive Distributed Lag (ARDL) model to global patent data from 1980 to 2010, we reveal that the innovation impact of different knowledge structures varies significantly across temporal dimensions. Specifically, KSD exerts a positive influence on disruptive innovation in the long run, affirming its role in enabling cross-boundary novelty and technological recombination. However, its short-run

effect is negative, reflecting the coordination burdens and integration frictions associated with heterogeneous knowledge inputs. KB shows a significant long-run negative effect, suggesting that excessive internal diversification may dilute technological coherence and hinder breakthrough potential. In contrast, KD contributes positively in the short run, but its long-run influence is not significant, highlighting the temporal limits of domain-specific specialization.

These findings offer practical insights into how innovation systems can reconcile the temporal trade-offs inherent in leveraging diverse knowledge structures. In particular, the short-term coordination burden and long-term disruptive potential of KSD underscore the need for governance structures that are explicitly designed to absorb temporal friction. Rather than merely increasing collaboration, innovation infrastructures must function as temporal bridges—buffering early-stage integration inefficiencies while preserving long-term recombinability. To achieve this, governments and funding agencies should support modular and phase-based knowledge integration mechanisms, such as two-stage public-private R&D consortia that separate exploratory knowledge matching from solution development phases. Additionally, platform-based digital infrastructure (e.g., centralized research asset registries, structured metadata repositories) can be developed to reduce search and alignment costs among disparate actors during early-stage collaboration. Regarding the long-run negative effects of KB, the results suggest that while internal interdisciplinarity holds conceptual appeal, it may introduce latent coordination complexity over time. Therefore, knowledge integration within single organizations should be governed through strategic modularization. Funding programs and institutional evaluations should move away from undirected interdisciplinarity and instead encourage bounded integration, such as matrix organizational structures that allow domain-specific subunits to recombine outputs selectively, avoiding wholesale internal diffusion. Furthermore, mid-term evaluation checkpoints can help prevent project over-extension by identifying when internal breadth begins to hinder coherence. Finally, the short-run positive but long-run insignificant role of KD highlights that short-term technical expertise alone is insufficient to sustain breakthrough trajectories. Policy frameworks should therefore incentivize depth-to-diversity transitions over time. For example, project funding could adopt tapered incentive schemes, in which early-stage funding rewards technical depth, while renewal or scaling-up depends on demonstrable cross-domain expansion. Additionally, career development tracks in public R&D institutions can be designed to encourage temporal diversification—starting from vertical expertise and gradually incorporating horizontal collaborations, ensuring that individual-level knowledge accumulation aligns with systemic innovation needs.

This study also has several limitations that warrant further investigation. First, our analysis adopts the CD index as the sole measure of disruptive innovation. While this indicator has been validated in recent large-scale studies, alternative metrics such as novelty scores, radicalness indicators, or paradigm-shift detection frameworks may capture different facets of disruption. Future research could explore the robustness of our results by substituting or triangulating CD with these alternative outcome measures. Second, although this study treats KB and KSD as independent dimensions, we acknowledge that their relationship may be more complex. In particular,

conceptual breadth may partially arise from exposure to diverse knowledge sources, suggesting potential endogeneity or interaction effects. Our current model specification does not explicitly test for such interdependencies. Future work could address this by introducing interaction terms, structural equation model, or dynamic panel techniques to capture potential co-evolution or causal links between KB and KSD over time.

Acknowledgments

References

Abdul Basit, S., & Medase, K. (2019). The diversity of knowledge sources and its impact on firm-level innovation: Evidence from Germany. European Journal of Innovation Management, 22(4), 681-714.

Alblooshi, M., Shamsuzzaman, M., & Haridy, S. (2021). The relationship between leadership styles and organisational innovation: A systematic literature review and narrative synthesis. European Journal of Innovation Management, 24(2), 338-370.

Boh, W. F., Evaristo, R., & Ouderkirk, A. (2014). Balancing breadth and depth of expertise for innovation: A 3M story. Research Policy, 43(2), 349-366.

Bowen III, D. E., Frésard, L., & Hoberg, G. (2023). Rapidly evolving technologies and startup exits. Management Science, 69(2), 940-967.

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. Journal of the Royal Statistical Society Series B: Statistical Methodology, 37(2), 149-163.

Chevalier‐Roignant, B., Flath, C. M., & Trigeorgis, L. (2019). Disruptive innovation, market entry and production flexibility in heterogeneous oligopoly. Production and Operations Management, 28(7), 1641-1657.

Christensen, C. M. (2015). The innovator's dilemma: when new technologies cause great firms to fail. Harvard Business Review Press.

Darroch, J. (2005). Knowledge management, innovation and firm performance. Journal of knowledge management, 9(3), 101-115.

Dogru, T., Mody, M., & Suess, C. (2019). Adding evidence to the debate: Quantifying Airbnb's disruptive impact on ten key hotel markets. Tourism Management, 72, 27-38.

Edmondson, A. C., & Mogelof, J. P. (2006). Explaining psychological safety in innovation teams: organizational culture, team dynamics, or personality? In Creativity and innovation in organizational teams (pp. 129-156). Psychology Press.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. Management science, 63(3), 791-817.

Grant, R. M. (1996). Toward a knowledge‐based theory of the firm. Strategic management journal, 17(S2), 109-122.

Hajialibeigi, M. (2023). Is more diverse always the better? External knowledge source clusters and innovation performance in Germany. Economics of Innovation and New Technology, 32(5), 663-681.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools.

Hang, C. C., Garnsey, E., & Ruan, Y. (2015). Opportunities for disruption. Technovation, 39, 83-93.

Harris, R., & Sollis, R. (2003). Applied Time Series Modelling and Forecasting.

Irfan, M., Razzaq, A., Sharif, A., & Yang, X. (2022). Influence mechanism between green finance and green innovation: exploring regional policy intervention effects in China. Technological Forecasting and Social Change, 182, 121882.
Jalil, A., & Ma, Y. (2008). Financial development and economic growth: Time series evidence from Pakistan and China. Journal of economic cooperation, 29(2), 29-68.
Jin, X., Wang, J., Chen, S., & Wang, T. (2015). A study of the relationship between the knowledge base and the innovation performance under the organizational slack regulating. Management Decision, 53(10), 2202-2225.
Kaplan, S., & Vakili, K. (2015). The double‐edged sword of recombination in breakthrough innovation. Strategic Management Journal, 36(10), 1435-1457.
Kim, H., Hwang, S. J., & Yoon, W. (2023). Industry cluster, organizational diversity, and innovation. International Journal of Innovation Studies, 7(3), 187-195.
Leiponen, A., & Helfat, C. E. (2010). Innovation objectives, knowledge sources, and the benefits of breadth. Strategic management journal, 31(2), 224-236.
Levina, M. (2017). Disrupt or die: Mobile health and disruptive innovation as body politics. Television & New Media, 18(6), 548-564.
Luo, T., Qu, J., & Cheng, S. (2024). Knowledge Network Embeddedness and Innovation Resilience. IEEE Transactions on Engineering Management.
Mardani, A., Nikoosokhan, S., Moradi, M., & Doustar, M. (2018). The relationship between knowledge management and innovation performance. The Journal of High Technology Management Research, 29(1), 12-26.
Mushtaq, R. (2011). Augmented dickey fuller test.
Nagy, D., Schuessler, J., & Dubinsky, A. (2016). Defining and identifying disruptive innovations. Industrial marketing management, 57, 119-126.
Nelson, R. R. (1985). An evolutionary theory of economic change. harvard university press.
Osiyevskyy, O., & Dewald, J. (2015). Explorative versus exploitative business model change: the cognitive antecedents of firm‐level responses to disruptive innovation. Strategic Entrepreneurship Journal, 9(1), 58-78.
Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. Nature, 613(7942), 138-144.
Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. Journal of applied econometrics, 16(3), 289-326.
Pesaran, M. H., Shin, Y., & Smith, R. P. (1999). Pooled mean group estimation of dynamic heterogeneous panels. Journal of the American statistical Association, 94(446), 621-634.
Reinhardt, R., & Gurtner, S. (2015). Differences between early adopters of disruptive and sustaining innovations. Journal of Business Research, 68(1), 137-145.
Rodriguez, M., Doloreux, D., & Shearmur, R. (2017). Variety in external knowledge sourcing and innovation novelty: Evidence from the KIBS sector in Spain. Technovation, 68, 35-43.
Ruan, Y., Hang, C. C., & Wang, Y. M. (2014). Governmentʹ s role in disruptive innovation and industry emergence: The case of the electric bike in China. Technovation, 34(12), 785-796.
Suseno, Y. (2018). Disruptive innovation and the creation of social capital in Indonesia's urban communities. Asia Pacific Business Review, 24(2), 174-195.
Thakur–Wernz, P., & Wernz, C. (2022). Impact of stronger intellectual property rights regime on innovation: Evidence from de alio versus de novo Indian bio-pharmaceutical firms. Journal of Business Research, 138, 457-473.
Van Balen, T., Tarakci, M., & Sood, A. (2019). Do disruptive visions pay off? The impact of disruptive entrepreneurial visions on venture funding. Journal of Management Studies, 56(2), 303-342.

Wen, J., Qualls, W. J., & Zeng, D. (2021). To explore or exploit: The influence of inter-firm R&D network diversity and structural holes on innovation outcomes. Technovation, 100, 102178.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036-1039.

Xu, S. (2015). Balancing the two knowledge dimensions in innovation efforts: an empirical examination among pharmaceutical firms. Journal of product innovation management, 32(4), 610-621.

Zhou, K. Z., & Li, C. B. (2012). How knowledge affects radical innovation: Knowledge base, market knowledge acquisition, and internal knowledge sharing. Strategic management journal, 33(9), 1090-1102.

Zou, B., Guo, F., & Guo, J. (2019). Antecedents and outcomes of breadth and depth of absorptive capacity: An empirical study. Journal of Management & Organization, 25(5), 764-782.

https://doi.org/10.51408/issi2025_008

Web Mining the Online Presence of Global Scientific Academies

Xiaoli Chen[1], Xuezhao Wang[2]

[1]*chenxl@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences (China)

[2]*Wangxz@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences (China)
University of Chinese Academy of Sciences (China)

Abstract

Global scientific academies have been adapting their role in fostering scientific communication and promoting science since their inception in the 15th century. Despite their prominence, the institutional norms and identities of scientific academies remain underexplored. In the digital age, their websites reflect their evolving roles, organizational priorities, and the balance between conformity and innovation. This study examines how scientific academies structure their online identities through content organization and communication strategies.

This study employs web mining techniques to analyze large-scale academy website data. It uncovers structural patterns and behavioral trends in how scientific academies present themselves online. Formal Concept Analysis (FCA) is applied to develop a unified taxonomy, enabling systematic comparisons of digital strategies across multiple academies. Using institutional theory, this study uses quantitative method to examine how academies balance conformity with differentiation in their digital presence. The research addresses two core questions: (RQ1) What content and communication patterns are adopted by global scientific academies in their online presence? And (RQ2) How do scientific academies balance imitation and innovation in their digital strategies?

The findings identify distinct website content patterns, showing how academies balance tradition and adaptation in their digital presence. Hierarchical clustering reveals three strategic approaches: (1) highly innovative academies that introduce novel digital structures, (2) conservative academies that show fragmented or underdeveloped structures, and (3) hybrid academies that combine imitation with selective innovation. The study also highlights key thematic differences in content emphasis, such as governance, scientific cooperation, and public outreach. These insights contribute to institutional theory and scholarly communication studies, revealing how scientific academies use their online presence to maintain legitimacy, engage the public, and foster international collaboration.

This study highlights common features of scientific academies' online presence, including an emphasis on membership, strategic planning, and scholarly communication to reinforce institutional legitimacy. Additionally, academies adapt their digital strategies to facilitate scientific collaboration in response to evolving societal expectations. Innovative activities include increasing transparency on the academy's decisions, achievements, budget, yearbooks, and interactive digital engagement strategies. These activities enhance public trust in scientific academies and science itself while improving communication efficiency. These findings offer guidance for scholars, academy leaders, and policymakers seeking to optimize digital engagement strategies and strengthen global scientific networks in the digital era.

Introduction

Scientific academies have long served as the cornerstone of knowledge advancement and scholarly communication since their inception in the 15th century. As technology advances and global interconnectivity increases, scientific academies

rely on their digital presence to extend influence, disseminate research, and engage with a diverse audience worldwide. Despite the increasing prominence of digital communication, little research examines how they structure their online presence and institutional identity. Additionally, there is limited understanding of how these academies balance imitation—adopting common practices—and innovation—developing unique digital strategies—in their approach to web-based communication.

Research on institutional digital presence has largely focused on universities (Lepori et al., 2014; Will & Callison, 2006), governmental organizations (Neumann et al., 2022), and research institutions (Burford, 2014; Elsayed, 2017), leaving scientific academies underexplored. Web mining has been applied to map innovation ecosystems (Kinne & Axenbeck, 2020), predict firm-level innovation (Axenbeck & Breithaupt, 2021; Kinne & Lenz, 2021), and analyze the accessibility of digital platforms (Singh et al., 2024; Alim, 2021). However, little research has specifically addressed the digital strategies of scientific academies. Unlike universities or firms, scientific academies operate at the intersection of academic prestige, policy influence, and public engagement, making their digital behavior distinct. This study utilizes prior web mining methodologies by analyzing how academies structure digital content, offering a comparative framework to assess the balance between imitation and innovation in the academies digital strategy.

This study is grounded in institutional theory, which provides a framework for understanding how scientific academies navigate community expectations and the tension between conformity and differentiation. Institutional theory explains how organizations conform to external expectations through institutional isomorphism. This process includes coercive pressures (regulatory and funding requirements), mimetic pressures (emulating successful peers), and normative pressures (adhering to professional standards and societal expectations). This theory framework provides an explanation on how scientific academies structure their online presence, influencing whether they conform to widely accepted digital taxonomies, adopt innovative approaches to distinguish themselves, or balance both strategies to maintain legitimacy while adapting to evolving scientific and societal demands. The web-based content strategies reflect their efforts to adhere to professional norms, align with stakeholder expectations, and assert their role as authoritative scientific institutions. At the same time, they face the challenge of distinguishing themselves through novel digital practices. This study uses quantitative method builds on institutional theory to analyze how scientific academies balance conformity and differentiation in their digital strategies.

To investigate these dynamics, this study utilizes web mining techniques combined with Formal Concept Analysis (FCA) to analyze the online presence of scientific academies. The hierarchical relations of web content are harvested by web mining techniques. FCA is employed to construct a unified taxonomy from the extracted hypernym-hyponym pairs. The unified taxonomy identifies patterns in content structure and content organization across these academies' websites. By quantitatively comparing these patterns, this study aims to uncover how academies engage with stakeholders, promote collaboration, and contribute to scientific

discourse on a global scale. The research is guided by two primary research questions: (RQ1) What content and communication patterns are adopted by global scientific academies in their online presence? And (RQ2) How do scientific academies balance imitation and innovation in their digital strategies?

This research makes several key contributions. First, it introduces a novel application of web mining and FCA to analyze how scientific academies structure their web content, providing a scalable and systematic method for web content taxonomy construction. Second, it advances institutional theory by exploring how academies cope with mimetic and normative pressures in shaping their digital strategies, which is reflected by their balance between imitation and innovation. Third, it provides practical insights for scholars, institutional leaders, and policymakers seeking to optimize digital engagement strategies. Understanding how academies structure their online presence can inform the development of more effective digital communication frameworks, enhance public engagement, and strengthen global scientific networks.

Related Work

The study of institutional digital strategies has gained increasing significance as institutions leverage digital platforms for communication, collaboration, and knowledge dissemination. While universities, government agencies, and firms have been extensively studied, scientific academies remain an overlooked category despite their critical role in shaping global scientific discourse. This research builds upon prior studies in institutional digital identity, web mining, and content taxonomy to assess how scientific academies structure their online presence.

Prior research has explored how institutions use digital platforms to shape institutional identity. Research has shown that institutional priorities shape online strategies across different organizations, including universities (Lepori et al., 2014; Will & Callison, 2006), government agencies (Neumann et al., 2022), and research institutions (Burford, 2014; Elsayed, 2017). Comparative studies on scientific academies (Isavand & Poormoghim, 2024) have examined regional differences but lack a global perspective on digital engagement strategies.

Studies in content organization and web architectures further demonstrate how institutions adapt their online presence to align with strategic goals (Campos et al., 2019; Karanasios et al., 2013). However, these studies primarily focus on universities and corporate entities, leaving a gap in understanding how scientific academies balance tradition and digital transformation.

Web mining has been widely applied in analyzing Organizational Structures and innovation behaviors. Researchers have used web data to map innovation ecosystems (Kinne & Axenbeck, 2020; Kinne & Lenz, 2021), predict firm-level digital strategies (Axenbeck & Breithaupt, 2021), and classify academic webpages (Kenekayoro et al., 2014, 2015). Historical web archives (Schroeder et al., 2020; Tsakalidis et al., 2021) provide insights into the evolution of institutional priorities, demonstrating how digital structures shift over time. Despite these advancements, scientific academies remain largely absent from web mining research, even though they play a crucial role in balancing scientific legitimacy, policy influence, and public

engagement. Prior methodologies have not been applied to systematically analyze how these institutions construct their digital presence.

The tension between institutional imitation and innovation is central to understanding how institutions adopt digital strategies. Institutional theory identifies coercive (regulatory), mimetic (peer-driven), and normative (professional) pressures as key factors shaping institutional behavior in digital spaces (Engelbrecht et al., 2020, 2022; Cox, 2007, 2008). Research on higher education institutions (Lepori et al., 2014) and corporate strategies (Gök et al., 2015; Thelwall, 2006) suggests that institutions often emulate established digital norms while attempting to differentiate themselves.

However, scientific academies face a unique challenge: upholding scientific authority and global credibility while adapting to national policy environments. Unlike universities, which primarily engage academic audiences, academies must also address policymakers, funding agencies, and the public. Prior research has not systematically examined how scientific academies navigate these competing demands in digital spaces. While previous studies have applied web mining, content classification, and institutional theory to universities, firms, and government agencies, no study has systematically examined the digital presence of scientific academies on a global scale. Unlike commercial enterprises, which optimize digital strategies for competitive advantage, scientific academies must balance scientific prestige, national policies, and public engagement. Furthermore, while studies on content classification and historical web evolution (Campos et al., 2019; Tsakalidis et al., 2021) provide foundational insights, they do not assess how scientific academies' digital strategies reflect their institutional missions.

This study builds on these research strands by integrating insights from web mining, institutional behavior, and content taxonomy to examine the digital presence of global scientific academies. This study addresses the current research gap by applying web mining and Formal Concept Analysis (FCA) to systematically examine how scientific academies structure their digital presence. A comparative framework is proposed for assessing how academies adapt to scientific norms and policy expectations in their online representations. By integrating insights from institutional theory, web mining, and web content taxonomies, this research advances our understanding of how scientific academies construct and maintain legitimacy in the digital age.

Methodology

This study applies web mining techniques and Formal Concept Analysis (FCA) to analyze systematically the digital presence of global scientific academies. It also explores institutional digital strategies to improve theoretical understanding in this area. This methodology addresses research questions by identifying patterns of imitation and innovation in the digital communication strategies of scientific academies. The methodology consists of five distinct phases, as illustrated in Figure 1: (1) Website Data Harvesting. Extracting structured information from global scientific academy websites. (2) Hierarchical Concept Development. Identifying hypernym-hyponym relationships to model content structures. (3) Category

Development and Category Mapping. Grouping content into meaningful categories using LLMs and word embeddings. (4) Taxonomy Construction. Applying FCA and graph pruning to refine hierarchical structures. (5) Comparative Analysis. Evaluating the thematic and structural commonalities and differences among academies.

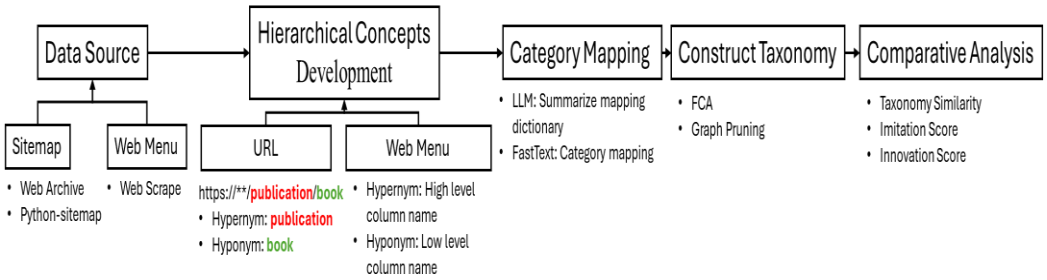


Figure 1. Framework for Web Mining and Comparative Analysis of Scientific Academies.

Website Data Harvesting

The website sitemaps provide a comprehensive structural blueprint of each academy's web presence, listing URLs that encapsulate both structural and content-related aspects. However, official sitemaps are sometimes incomplete or unavailable. The primary source of data is the sitemaps of scientific academies that are retrieved from the website archival platform[1]. As a complementary approach, automated sitemap generators like *python-sitemap*[2] are used to reclaim web pages that may be missing. However, both methods may encounter challenges due to scraping restrictions and web connection issues. To mitigate these limitations, this study uses navigation menus of scientific academies' websites to supplement data collection. Compared to sitemaps, website navigation menus provide another perspective on content organization. These menus typically highlight the key focus of institutional priorities and mission. However, this method has limitations—some websites lack a well-structured navigation website menu or offer shallow categorizations. This study combines the three data sources, and a manual check by random browsing of the website is also conducted to verify the key aspects of the website's columns are included in the data collection.

Hierarchical Concept Development

This study analyzes the hierarchical structure of website content from global scientific academies. It extracts hypernym-hyponym relationships from menu items and webpage URLs. Each URL is stripped of its domain and segmented hierarchically using the forward-slash (/) delimiter, as illustrated in Figure 1. The resulting hierarchical dictionary preserves the hypernym-hyponym relationships within the website's navigation content and webpage URLs, with higher-level menu

[1] https://web.archive.org/
[2] https://github.com/c4software/python-sitemap

items or URL relative paths (hypernyms) containing more specific subcategories (hyponyms).

To ensure cross-institutional consistency of scientific academies, this study utilizes the DeepSeek[3] Large Language Model (LLM) for language translation, normalizing terminologies across different linguistic contexts. Given that the hypernym and hyponym pairs extracted from the URL path often include acronyms and numbers, this study WordNet to retain only semantically meaningful terms. This process refines the extracted relationships and improves taxonomy accuracy.

> *LLM Instructions:*
> *You are an expert in hierarchical content classification and taxonomy development. Your task is to refine a set of extracted hypernym-hyponym pairs by identifying meaningful concepts and filtering out irrelevant terms.*
> *1. Input Format: You will receive a list of hypernym-hyponym pairs extracted from website structures.*
> *2. Objective: Identify core concepts by:*
> *- Grouping similar hyponyms under a meaningful hypernym.*
> *- Removing noisy terms, such as acronyms, numbers, and ambiguous words.*
> *- Ensuring logical consistency in hierarchical relationships.*
> *3. Output Format:*
> *- A structured JSON object where each hypernym maps to refined hyponyms.*

Categories Development and Category Mapping

After cleansing the extracted hypernym-hyponym pairs, this study establishes core concepts that form the foundation of the taxonomy. This process involves Identifying and summarizing content patterns using an LLM pipeline. These pattern words filter out irrelevant hypernyms and hyponyms to enhance dataset clarity.

For computational efficiency, the FastText model is used to compute the average word embeddings of hypernym and hyponym terms. Cosine similarity scores of these embedding vectors are mapped into the nearest normalized category embedding. To maintain classification integrity, manual verification is conducted, resolving inconsistencies and improving accuracy.

Developing Website Content Taxonomy

This study applies Formal Concept Analysis (FCA) to structure and refine hierarchical web content. FCA constructs a concept lattice, while graph pruning enhances consistency, reduces redundancy, and optimizes hierarchical relationships. Formal Concept Analysis is a well-established method for knowledge organization. It is particularly suited for this task as it enables the construction of a concept lattice,

[3] https://www.deepseek.com/

effectively capturing relationships between categories while preserving the hierarchical nature of web structures. In this study, FCA is applied to generate a formal taxonomy of web content from scientific academy websites, facilitating comparative analysis.

The formal context is represented as a binary relation $K = (G, M, I)$, where:

- Objects (G): $g_i \in G$ denotes hyponyms (specific subcategories in the taxonomy).
- Attributes(M): $m_j \in M$ denotes hypernyms (general categories representing broader concepts).
- Incidence Relation (I): A binary relation $I \subseteq G \times M$ indicating which objects belong to which attributes. The relation is represented as a binary matrix B, where:

$$B_{ij} = \begin{cases} 1, & \text{if object } g_i \text{ is associated with attribute } m_i \\ 0, & otherwise \end{cases}$$

Using this matrix representation, the attribute derivation operator A' and the object derivation operator B' could be derived:

$$A' = \{m \in M \mid \forall g \in A, (g, m) \in I\}$$
$$B' = \{m \in M \mid \forall g \in B, (g, m) \in I\}$$

Here A' is the set of all attributes shared by objects in A. B' is the set of all objects sharing the attributes in B. A formal concept is a pair (A, B), where:

$$A = B' \text{ and } B = A'$$

Here A is the extent, which means the set of all objects (hyponyms) belonging to concept B. B is the intent, which means the set of all attributes (hypernyms) that describe all objects in A. A concept lattice $L(K)$ is formed by structuring these concepts into a partially ordered set:

$$(A_1, B_1) \leq (A_2, B_2)\ if\ A_1 \subseteq A_2\ (or\ equivalently\ B_2 \subseteq B_1)$$

This implies that more general concepts are ranked higher in the lattice, while specific concepts appear lower. This is implemented by using Meet ($\wedge$) operation and Join ($\vee$). Meet ($\wedge$) operation computes the greatest lower bound of two concepts, used to identify hyponym terms:

$$(A_1, B_1) \wedge (A_2, B_2) = (A_1 \cap A_2, (A_1 \cap A_2)')$$

Join ($\vee$) operation computes the greatest least upper bound of two concepts, used to identify hypernym terms:

$$(A_1, B_1) \vee (A_2, B_2) = (B_1 \cap B_2, (B_1 \cap B_2)')$$

A key challenge of FCA is multi-parent assignments, where a single hyponym is linked to multiple hypernyms, potentially creating ambiguous or cyclic relationships. Due to the diverse structures of academy websites, the extracted categories often exhibit inconsistent terminology, redundancies, and overlapping concepts. To further

refine the extracted hierarchical taxonomy, this study applies a graph pruning method to ensure hierarchical consistency, eliminate conflicts, and resolve structural inconsistencies.

To effectively resolve cyclic dependencies in hypernym-hyponym pairs, conflict cycles were detected using depth-first search (DFS). Manual evaluation was then conducted to eliminate incorrect hypernym-hyponym relationships while retaining only the most contextually appropriate ones. Multi-parent issues were addressed using a similar manual resolution process. For example, if the term "*Funding*" appeared as a subcategory under both "*Governance*" and "*Supporting Science*", the pruning process ensured its placement under "*Supporting Science*", where it aligns with funding mechanisms for scientific projects rather than administrative governance. Additionally, cyclic dependencies—such as a category incorrectly appearing as both a parent and a child (e.g., "Awards" categorized under both " *Supporting Science*" and "*Knowledge Resources*")—were detected using depth-first search (DFS) and manually resolved to preserve logical consistency in the taxonomy. The iterative manual review ensured that meaningful hierarchical relationships were maintained, preventing redundancy and ambiguity. To validate the accuracy of the final taxonomy, a manual review is conducted for a subset of academy websites, ensuring that the taxonomy aligns with real-world institutional practices.

Comparative Analysis of Global Scientific Academies

This study leverages the constructed taxonomy to examining thematic and structural differences in their digital presence. One aspect of comparison is assessing the overall scale of the websites, including the number of pages they contain, to gauge their digital footprint. Levels of URL paths are analyzed to understand how deep the content structure is, which reflects the complexity and organization of the websites. Analyzing the balance between imitation and innovation in website structures is crucial for understanding how scientific academies establish their digital presence. This study develops a methodology based on a combination of similarity analysis and unique content evaluation to quantify the extent to which websites adopt existing taxonomies, imitate peer's digital practice and introduce novel structures. Each website's hypernym-hyponym pairs were enriched by identifying and incorporating missing parent nodes from the common taxonomy to ensure structural completeness. For each site *s*, similarity to common taxonomy is assessed how closely each website adhered to the common taxonomy by computing its cosine similarity to the taxonomy. It is a balance of how many of the site's hypernym-hyponym pairs are present in the common taxonomy (Precision(s)) and how much of the taxonomy is covered by the site (Recall(s)). For site *s,* where $P_s = \{(h_k, h'_k) \mid h_k$ is a hypernym of $h'_k\}$ is the set of hypernym-hyponym pairs of site *s*. $P_{taxonomy}$ is the set of hypernym-hyponym pairs of a common taxonomy. The similarity analysis is conducted to inspect each website's similarity with the common taxonomy by performing the following method:

$$Taxonomy\ Similarity(s) = \frac{2 \times Precision(s) \times Recall(s)}{Precision(s) + Recall(s)}$$

where $Precision(s) = \frac{/P_s \cap P_{\text{taxonomy}}/}{/P_s/}$ and $Recall(s) = \frac{/P_s \cap P_{\text{taxonomy}}/}{/P_{\text{taxonomy}}/}$.

To quantify the conformity between websites, this study introduces the Imitation Score based on the average similarity to other websites. Each site *s* is represented as a binary vector v_s of length *d*, where d is the total number of unique hypernym-hyponym pairs across all websites. These pairs align with the taxonomy structure. Each entry in v_s is 1 if the corresponding hypernym-hyponym pair appears in the website, and 0 otherwise. Cosine similarity between two websites s_i and s_j is

$$cosine_sim(s_i, s_j) = \frac{v_{s_i} \cdot v_{s_j}}{|v_{s_i}||v_{s_j}|}$$

The imitation score for website *s* is its average cosine similarity with all other websites

$$Imitation\ Score(s_i) = \frac{1}{N-1} \sum_{s' \in S, s' \neq s} cosine_sim(s, s')$$

where *N* is the total number of websites.

To measure the uniqueness of a website's structure, this study computes an Innovation Score by comparing its hypernym-hyponym pairs with those of other websites. The Innovation Score for a website s is defined as the average number of unique hypernym-hyponym pairs it has compared to all other websites. Each website s_i is represented as a set of hypernym-hyponym pairs P_{s_i}. The uniqueness of s_i is determined by counting the number of pairs that do not exist in any other website s_j, where $j \neq i$.

$$Innovation\ Score(s_i) = \frac{1}{N-1} \sum_{s' \in S, s' \neq s} \left| P_{s_i} \setminus P_{s_j} \right|$$

Where $P_{s_i} = \{(h_k, h'_k) \mid h_k \text{ is a hypernym of } h'_k\}$ is a set of hypernym-hyponym pairs of website s_i. $P_{s_i} \setminus P_{s_j}$ denotes the set difference, capturing pairs that exist in s_i but not in s_j. The summation iterates over all other websites s_j, averaging the unique pairs. To ensure comparability between the Imitation Scores and the Innovation Scores, this study applies Min-Max Scaling for both the Imitation Scores and the Innovation Scores.

To further explore how academy websites differentiation, this study introduces a Distinctiveness Score to identify the most unique hypernym-hyponym pairs in each cluster. Given a set of clusters $C = \{C_1, C_2, \cdots, C_m\}$, each website s_i is assigned to a cluster c_j through hierarchical clustering:

$$f: S \to C, \quad f(s_i) = c_j$$

For each cluster c_j, this study aggregates all pairs from its member websites

$$P_{c_j} = \bigcup_{s_i \in c_j} P_{s_i}$$

The cluster-level frequency of a pair *(h, h')* is computed as

$$\text{count}_{c_j}(h, h') = \sum_{s_i \in c_j} 1\left((h, h') \in P_{s_i}\right)$$

The global frequency of a pair across all websites is:

$$\text{global_count}(h, h') = \sum_{s_i \in S} 1\left((h, h') \in P_{s_i}\right)$$

The relative frequency of a pair *(h, h')* in cluster c_j is given by

$$\text{relative_freq}_{c_j}(h, h') = \frac{\text{count}_{c_j}(h, h')}{\sum_{(h,h') \in P_{c_j}} \text{count}_{c_j}(h, h')}$$

The global probability of a pair appearing in the entire dataset is

$$P(h, h') = \frac{\text{global_count}(h, h')}{\sum_{(h,h') \in P} \text{global_count}(h, h')}$$

The distinctiveness score of a pair *(h, h')* in cluster c_j is

$$\text{distinctiveness}_{c_j}(h, h') = \frac{\text{relative_freq}_{c_j}(h, h')}{P(h, h')}$$

Statistical methods were used to validate the effectiveness of the cluster partition by distinguishing the imitation score and innovation score. Statistical methods were also applied to test if the identified the most distinctive hypernym-hyponyms and the least distinctive hypernym-hyponyms are significant in different types of scientific academies.

This study integrates Formal Concept Analysis (FCA), graph pruning, and manual verification to construct a reliable and accurate taxonomy, serving as the knowledge backbone for understanding the website content of scientific academies. To quantitatively assess digital strategies, Taxonomy Similarity, Imitation Score, and Innovation Score were developed to measure the extent to which academies adopt common practices, conform to established norms, and differentiate their digital presence. Additionally, the Distinctiveness Score was introduced to identify both the most unique and the most standardized content, providing insights into the balance between conformity and differentiation in the digital strategies of scientific academies.

Result and Analysis

The results of this study are organized into three main sections. The first section, Data Description, provides an overview of the dataset, detailing the structural and institutional patterns of scientific academy websites. The second section, Taxonomy of Scientific Academies' Web Content Organization, presents the taxonomy derived from Formal Concept Analysis (FCA) and graph pruning, demonstrating how these academies define their digital identities. The final section, Comparative Analysis of Digital Presence Across Scientific Academies, explores the balance between

imitation and innovation, revealing how different academies strategically position themselves within the global scientific community.

Data Description

This study utilizes the dataset of global scientific academies (Chen, 2024), focusing on a subset of 112 national scientific academies dedicated to the natural sciences and excluding those centered on medical and engineering disciplines. The sitemap and navigation menu data spans June to August 2024. After parsing and cleaning the datasets, and removing duplicate webpage entries and external links, 13,122,124 URLs from the sitemaps and 9,953 URLs from the navigation menus were retained for further analysis. These URLs were then analyzed using the taxonomy induction method outlined in the methodology section, which incorporates Formal Concept Analysis (FCA) and graph pruning. Through this process, 2,781 hypernym-hyponym pairs across the 112 websites were identified for content exploration and comparative analysis.

The analysis of the 112 academies reveals significant variation in the size and organization of their web content. The number of URLs in the sitemaps varies widely, ranging from 30 to 1.5 million, with an average of 70,000 URLs per academy. Similarly, the number of items in the navigation menus ranges from 3 to 211, with an average of 40 menu items per academy. These variations indicate differing digital strategies, where some academies maintain extensive online repositories, while others prioritize streamlined, high-level navigation.

Figure 2 visualizes the depth distribution of URLs across different academies, mapping the relationship between the total number of URLs and their hierarchical depth. This analysis shows that academies with larger numbers of URLs do not necessarily structure their content deeper within the hierarchy. The lack of significant correlation, confirmed by a linear correlation analysis (p-value = 0.334), suggests that different content organization strategies influence website structure beyond mere scale. Some academies may prioritize broad, shallow hierarchies for accessibility, while others adopt deeper structures for detailed content segmentation.

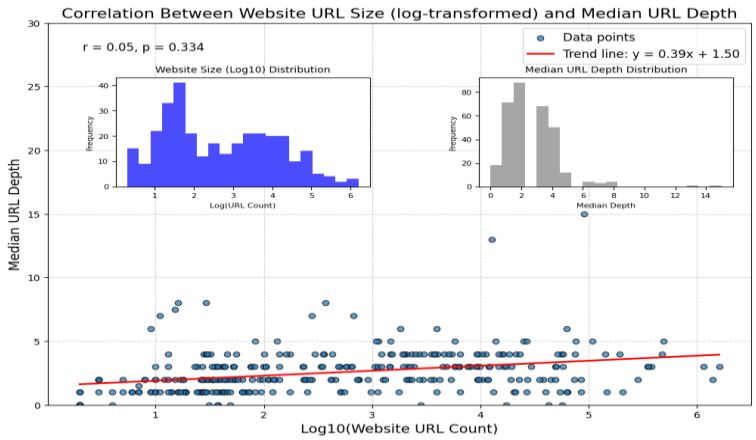


Figure 2. Correlation Between Website URL Size (Log-transformed) and Median URL Depth.

Taxonomy of Web Content

This study applies Formal Concept Analysis (FCA) and graph pruning to develop a structured taxonomy for global scientific academies' web content. The resulting classification identifies 121 unique hypernym-hyponym pairs (Figure 3). The primary categories identified through FCA are "*Governance*", "*Information*", "*Knowledge Resources*", "*Organizational Role*", "*Organizational Structure*", "*Public Outreach*", "*Scientific Cooperation*", and "*Supporting Science*". Each of the categories is further subdivided into specific subcategories that reflect the various areas of activities within these academies. These categories illuminate the institutional functions and strategic priorities of scientific academies, affirming their distinct yet overlapping roles in knowledge production, dissemination, and societal engagement. The taxonomy reveals three dominant functional categories—governance (as Learned Society archetype), public engagement (as Adviser to Society archetype), and scientific production (as Manager of Research archetype). These align with Engelbrecht et al.'s (2020) archetypes, demonstrating how academies balance internal organization, public engagement, and research leadership.

The Learned Society archetype is characterized by scientific academies as self-governing communities dedicated to fostering intellectual exchange and the advancement of knowledge. The taxonomy highlights the dominant presence of "*Organizational Structure*," "*Organizational Role*," and "*Governance*." These categories define the framework that supports scientific discourse and knowledge circulation. The legitimacy of learned societies is grounded in their ability to curate, manage, and disseminate scientific knowledge, a role further reinforced by their commitment to research documentation and public engagement.

The Adviser to Society archetype is evident in the emphasis on "*Public Outreach*", particularly in "*Science Communication*" and "*Science Advice*." These functions position scientific academies as intermediaries between researchers and the broader society. The findings suggest that academies use digital platforms to enhance scientific literacy, influence public understanding, and provide guidance on policy matters. The prominence of "*Knowledge Resources*" within this category underscores the dual responsibility of academies to engage both scientific professionals and the general public in knowledge exchange.

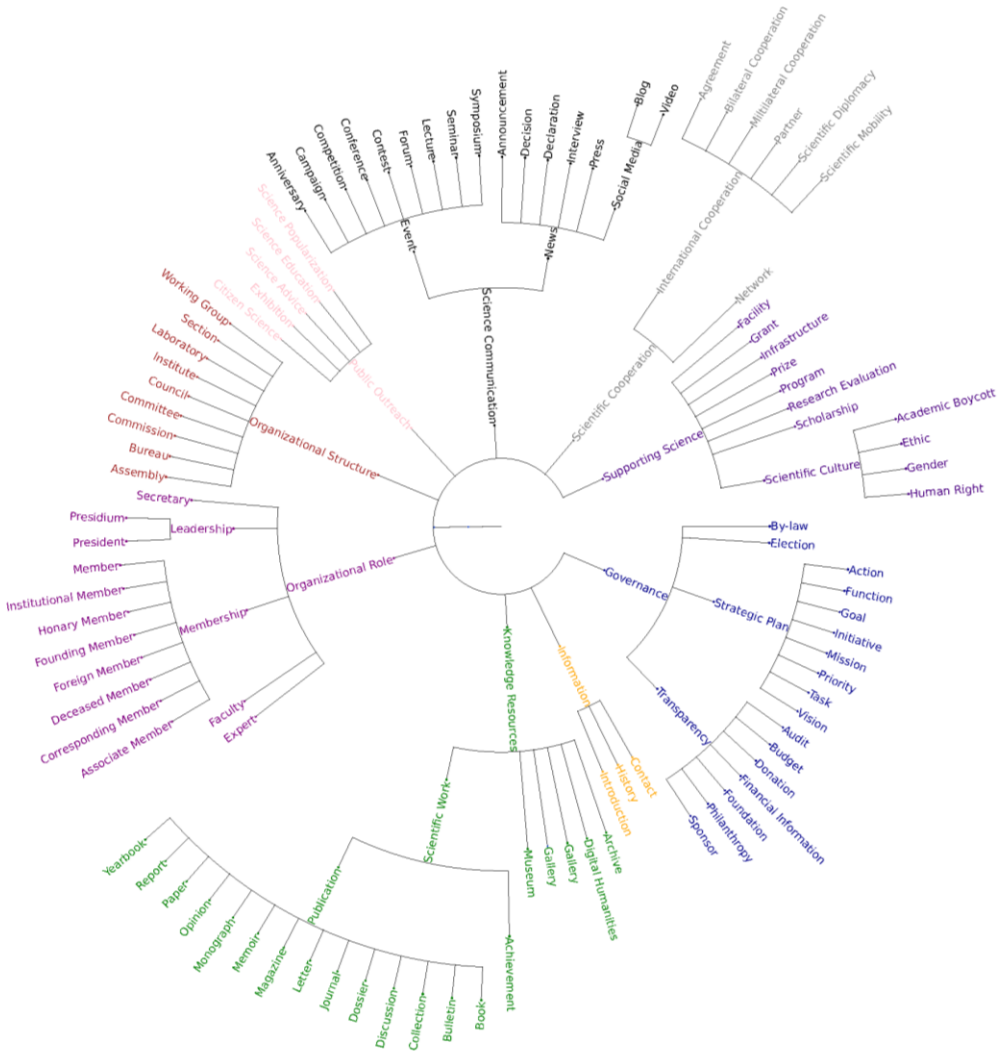


Figure 3. Hierarchical Taxonomy of Scientific Academies' Web Content.

The Manager of Research archetype extends beyond the direct management of research institutions to encompass a broader role in knowledge production and scientific excellence. The taxonomy demonstrates that "*Supporting Science*" and "*Scientific Cooperation*" are central to academy functions, signifying a strategic effort to cultivate both national and international scientific collaborations. The inclusion of "*Institution*" within the "*Organizational Structure*" of scientific academies suggests their direct involvement in knowledge creation.

Although Engelbrecht et al. (2020) primarily associated this archetype with direct research management, the taxonomy reveals that academies engage in a continuum of activities from knowledge production to dissemination. The presence of "*Knowledge Resources*" as a dominant category further illustrates that academies not only facilitate scientific research but also actively curate and preserve it. Some academies emphasize scientific recognition through awards and prizes, reinforcing

their role in advancing scientific excellence. The subcategory "*Archive*" within "*Knowledge Resources*" further highlights efforts to document and preserve national scientific and cultural heritage, reflecting a long-term commitment to maintaining and disseminating scientific knowledge.

The digital presence of global scientific academies is strategically structured to reflect their core missions and institutional priorities. The common taxonomy of these academies' websites reveals clear hierarchical relationships between key content categories, illustrating how they construct their institutional identity. The taxonomy highlights their role in facilitating knowledge circulation and maximizing its impact. It also identifies opportunities for public outreach, engagement, and independent advisory functions to governments. Scientific academies position themselves within a complex landscape of national and international policy, societal expectations, and intellectual networks, navigating challenges such as technological and resource disparities across institutions. Most academies do not fit neatly into a single archetype; even those within the same category may adopt distinct strategies to advance scientific excellence and promote public understanding of science. The following section will further examine the commonalities and unique characteristics of these academies' digital strategies.

Comparative Analysis of Digital Presence Across scientific academies

While global scientific academies share a common commitment to advancing scientific knowledge and assimilation knowledge, their online presences vary considerably. The Taxonomy Similarity score for the 112 scientific academies ranges from 0.2 to 0.75, with an average value of 0.42. This variation indicates differing degrees of alignment with the taxonomy developed. While some scientific academies closely follow the established taxonomy, others diverge in various ways, reflecting their unique priorities, missions, and regional contexts.

To better understand these variations, this study conducted a pairwise comparison of websites, utilizing hierarchical clustering based on Jaccard Similarity. This analysis revealed distinct groups of websites exhibiting different patterns in their hypernym-hyponym relationships. The dendrogram (tree diagram) in Figure 4 partitions the websites into three clusters, illustrating the degree of academies' web content similarity in structuring and categorization.

Figure 4 provides key insights into imitation and innovation behaviors across clusters. The left box plot in Figure 4 represents the Imitation Score for each cluster, which measures the average similarity of each website to others. Cluster 2 (green) has the lowest imitation score, meaning these websites are more unique and less similar to established patterns. This suggests a departure from conventional digital structures, possibly due to resource-limited context or underdeveloped website taxonomies. Cluster 1 (yellow) and Cluster 3 (red) have higher imitation scores, indicating stronger alignment with established conventions, implying that these websites adhere more closely to widely accepted content organization strategies.

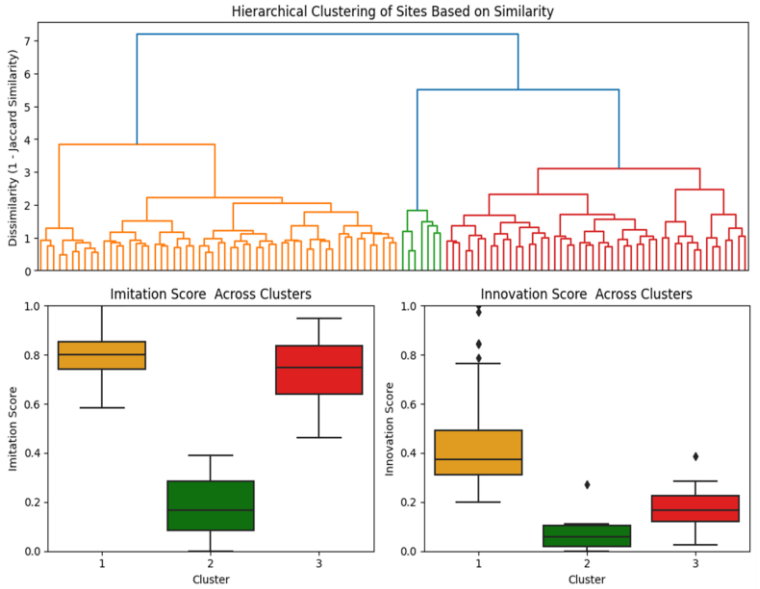


Figure 4. Hierarchical Clustering of Sites and Innovation/Imitation Scores of Website Groups.

Table 1. Descriptive Statistics for Taxonomy Similarities, Imitation Scores and Innovation Scores Across Clusters.

	Taxonomy Similarity			*Imitation Score*			*Innovation Score*		
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 1*	*Cluster 2*	*Cluster 3*
Value	▲ High (0.51)	▼ Low (0.21)	▼ Low (0.34)	▲ High (0.80)	▼ Low (0.18)	▲ High (0.73)	▲ High (0.44)	▼ Low (0.08)	▼ Low (0.17)
Mean	0.51	0.21	0.34	0.80	0.18	0.73	0.44	0.08	0.17
Median	0.48	0.19	0.35	0.80	0.17	0.75	0.37	0.06	0.17
SD	0.09	0.07	0.06	0.10	0.15	0.13	0.19	0.09	0.07
Min	0.39	0.15	0.22	0.58	0.00	0.46	0.20	0.00	0.03
Max	0.75	0.34	0.48	1.00	0.39	0.95	1.00	0.27	0.39
Cluster size	57	7	48	57	7	48	57	7	48
Overall Avg		0.35			0.57			0.23	
ANOVA F-Statistic		94.19**			88.33**			51.93**	
Cluster 1	1			1			1		
Cluster 2	10.63**	1		10.57**	1		8.14	1	
Cluster 3	11.89**	4.79**	1	2.96**	9.19**	1	9.81**	2.33**	1

Note: *p<0.05, **p<0.01.

The ANOVA test results for Taxonomy Similarity, Imitation Score and the Innovation Score have p-value <0.01, indicating highly significant difference in the three metrics across clusters. The Pairwise t-tests results of p-values <0.01 indicate clusters have distinct imitation and innovation behaviors. The Innovation Score of Cluster 2 and Cluster 3 do not show significance with p-value above 0.05. Bootstrap resampling is conducted before statistical analysis for robustness due to small sample sizes.

The right box plot in Figure 4 presents the Innovation Score, which captures the extent to which websites introduce new hypernym-hyponym relationships. Cluster 1 (yellow) has the highest innovation score, meaning that websites in this cluster introduce more unique content structures, signifying efforts toward digital differentiation. Cluster 2 (green) has the lowest innovation score, confirming that these websites not only diverge from common patterns but also lack substantial new actions. Cluster 3 (red) demonstrates moderate innovation, balancing between adopting conventional taxonomies and integrating some novel elements. Some academies adhere closely to established frameworks, while others diverge significantly. This divergence occurs either through the introduction of new structures or fragmented content strategies.

Table 1 summarizes the Taxonomy Similarity, the Imitation Scores, and the Innovation Scores across the three identified clusters. These statistics highlight how websites align with common taxonomies, maintain structural consistency, and introduce unique elements.

Cluster 1 (yellow) websites in Table 1 exhibit high innovation but low imitation scores, indicating that they are highly innovative academies that introduce novel digital structures. This suggests that these academies take a more innovative and forward-thinking approach to structuring their digital presence. Cluster 2 (green) websites have the lowest imitation and innovation scores, reflecting conservative digital strategies. These academies exhibit fragmented or underdeveloped web structures, often lacking clear content hierarchies or comprehensive navigation systems. This pattern may reflect a lack of cohesive content strategy, potentially hindering user navigation and information retrieval. Cluster 3 (red) websites show high imitation but low-to-moderate innovation, meaning they are hybrid academies that combine imitation with selective innovation. These websites prioritize standardization, ensuring consistency in their digital frameworks while making incremental refinements.

The statistical test results confirm that the clustering approach successfully identifies meaningful distinctions. ANOVA test results for Taxonomy Similarity, Imitation Score, and Innovation Score show p-values < 0.05, indicating statistically significant differences across clusters. Pairwise t-tests further validate distinct imitation and innovation behaviors across most clusters, except for a less significant difference in innovation behaviors between Cluster 2 and Cluster 3. Bootstrap resampling is applied to enhance robustness given varying sample sizes. These statistical findings reinforce the validity of the identified clusters and their implications for digital taxonomy structuring.

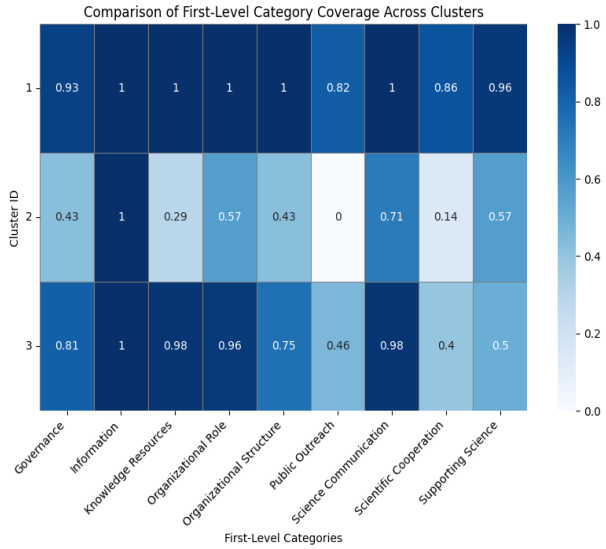


Figure 5. Comparison of First-Level Category Coverage Across Clusters.

Table 2. Distinctiveness and Statistical Analysis of Top Distinctiveness Pairs and Least Distinctiveness Pairs (Common Pairs) Across Clusters.

	Cluster 1			Cluster 2			Cluster 3		
	Distinct 5 Pairs	Other Pairs	Common 5 Pairs	Distinct 5 Pairs	Other Pairs	Common 5 Pairs	Distinct 5 Pairs	Other Pairs	Common 5 Pairs
Mean	1.54	1.49	0.54	11.11	1.03	0.47	2.19	1.09	0.28
Median	1.54	1.46	0.64	5.22	1.04	0.53	2.21	1.09	0.28
Std Dev	0.00	0.63	0.23	14.24	0.38	0.11	0.60	0.15	0.02
Min	1.54	0.50	0.22	3.65	0.43	0.35	1.55	0.74	0.26
Max	1.54	2.93	0.77	36.53	2.22	0.59	3.09	1.30	0.31
Hypernym-Hyponym Pairs Count	5	43	5	5	76	5	5	108	5
Distinct 5 Pairs	1			1			1		
Other Pairs	17.43**	1		1.47	1		4.93	1	
Common 5 Pairs	9.95**	-6.01**	1	1.67**	-8.66**	1	7.15**	-15.66**	1
Distinct 5 Pairs	Publication->Yearbook News->Decision Scientific Work->Achievement			Transparency->Audit Strategic Plan->Vision			Membership->Institutional Member Membership->Associate Member		

	Transparency->Budget Knowledge Resources->Museum	Membership->Corresponding Member Transparency->Financial Information Membership->Founding Member	Membership->Corresponding Member Membership->Honary Member Social Media->Blog
Common 5 Pairs	Membership->Corresponding Member Membership->Associate Member Strategic Plan->Vision Membership->Honary Member Membership->Founding Member	Knowledge Resources->Scientific Work Scientific Work->Publication Homepage->Scientific Cooperation News->Social Media Scientific Cooperation->International Cooperation	Supporting Science->Scholarship Event->Anniversary Event->Competition Publication->Memoir Organizational Structure->Assembly

Note: *p<0.05, **p<0.01.
The Pairwise t-tests results of p-values <0.01 indicate clusters have distinct imitation and innovation behaviors. The Distinct 5 Pairs and Other Pairs of Cluster 2 and Cluster 3 do not show significance with p-value above 0.05. Bootstrap Resampling is added before statistical analysis for robustness due to small sample sizes.

To gain deeper insights into how different websites cover the taxonomy categories, this study generated a heatmap (Figure 5) that visualizes the coverage of first-level categories across the 112 websites. The heatmap allows decision-makers to identify strengths and gaps in content representation. Academies can use this insight to align their digital strategies with common best practices while addressing areas of weak representation.

Cluster 1 (yellow) in Figure 5 exhibits the most comprehensive coverage across all first-level categories, with most values close to 1. Websites in this cluster consistently represent key categories, including "*Information*," "*Knowledge Resources*," "*Organizational Role*," "*Organizational Structure*," and "*Scientific Cooperation.*" This suggests that these websites follow a structured taxonomy, ensuring well-organized content and accessibility.

Cluster 2 (green) shows uneven category coverage, with "*Public Outreach*" (0.00) and "*Scientific Cooperation*" (0.14) largely absent, while "*Information*" (1.00) and "*Science Communication*" (0.71) are strongly represented. This suggests a selective emphasis on specific themes. This suggests that websites in this cluster focus on specific categories while omitting others, potentially indicating specialized or fragmented digital structures that reflect varied institutional priorities.

Cluster 3 (red) balances coverage, with high representation in "*Knowledge Resources*" (0.98), "*Organizational Structure*" (0.96), and "*Scientific Cooperation*" (0.98), while "*Public Outreach*" (0.46) and "*Supporting Science*" (0.50) are less prominent. This pattern suggests that websites in Cluster 3 generally align with common taxonomies but selectively emphasize certain content areas, striking a balance between conformity and differentiation.

These results confirm that clustering effectively differentiates websites based on their structural emphasis, highlighting distinct patterns in how scientific academies structure their online presence and the prioritization of content categories.

To further explore how specific content distinguishes scientific academies, this study applied the Distinctiveness Score (as outlined in the methodology section) to identify the most distinguishing hypernym-hyponym pairs and the least distinguishing hypernym-hyponym pairs. Table 2 presents a statistical analysis of the distinctiveness of hypernym-hyponym pairs across the three clusters, offering insights into differences in how websites structure their taxonomies. Cluster 2 exhibits the most unique structural elements, as indicated by its highest distinctiveness score (11.11) and high standard deviation (14.24). This suggests that websites in this cluster introduce the most unique structural elements. In contrast, Cluster 1 and Cluster 3 display lower distinctiveness scores (1.54 and 2.19, respectively), indicating a more moderate level of structural differentiation and stronger alignment with widely recognized taxonomies. The common pairs have significantly lower scores across all clusters (ranging from 0.28 to 0.54), confirming that frequently shared relationships follow more standardized patterns.

These findings highlight that while some academies maintain highly conventional taxonomies, others develop distinctive content structures, reflecting diverse institutional priorities and digital strategies. Pairwise t-tests confirm statistically significant differences ($p < 0.01$) between distinct and common pairs in Clusters 1

and 3, reinforcing clear structural separation. However, Cluster 2 and Cluster 3 do not show significant differences in "Other Pairs," indicating some shared taxonomy structures. These findings confirm that Cluster 2 exhibits the most structurally unique websites, while Cluster 1 and Cluster 3 balance imitation and innovation differently.
Distinct hypernym-hyponym pairs reveal unique digital strategies among scientific academies. Cluster 1 (Yellow) focuses on institutional knowledge, governance, and decision-making, emphasizing categories like "*Publication → Yearbook*" and "*Scientific Work → Achievement*" to document scholarly contributions. Cluster 2 (Green) emphasizes financial transparency and strategic vision, with categories like "*Transparency → Audit*" and "*Strategic Plan → Vision*," reflecting a focus on governance and long-term planning. Cluster 3 (Red) prioritizes digital engagement, using categories like "*Social Media → Blog*" and "*Membership → Honorary Member*" to create an interactive outreach strategy.
These distinctions illustrate how different academies adapt their digital presence based on governance models, transparency requirements, and audience engagement strategies. Common hypernym-hyponym pairs highlight shared digital structures across scientific academies. Most emphasize structured membership systems, with categories like "*Membership → Corresponding Member / Associate Member / Honorary Member / Founding Member*," reinforcing their role as academic communities. Strategic foresight and institutional direction remain central, evidenced by "*Strategic Plan → Vision*." Scientific cooperation and public communication are common priorities. Categories like "*Scientific Cooperation → International Cooperation*" and "*News → Social Media*" demonstrate the widespread use of digital platforms for knowledge dissemination and stakeholder engagement.

Discussion

The findings of this study highlight both shared and divergent patterns in how global scientific academies structure their online presence. Addressing RQ1, the taxonomy derived from Formal Concept Analysis (FCA) reveals a common framework that organizes academy websites around governance, knowledge resources, public outreach, scientific cooperation, and organizational structures. Despite this shared foundation, academies vary in how they emphasize these elements. Some prioritize structured governance and scholarly documentation, while others focus on enhancing public outreach or fostering scientific collaborations. These differences reflect the diverse roles academies play in their national and international contexts, shaping how they present their digital identities.
For RQ2, the comparative analysis of taxonomy similarity, imitation scores, and innovation scores demonstrates varying levels of adherence to standard digital frameworks. Academies in Cluster 1 exhibit high innovation but low imitation scores, indicating that they are highly innovative academies that introduce novel digital structures. In contrast, those in Cluster 2 show the lowest imitation and innovation scores, characterized by fragmented or underdeveloped digital strategies that suggest conservative digital strategies. Cluster 3 aligns closely with established taxonomies,

maintaining consistency while integrating selective innovations. These variations underscore how scientific academies navigate the balance between digital conformity and differentiation. The most distinctive hypernym-hyponym pairs reveal areas where academies differentiate themselves, such as financial transparency initiatives or interactive digital engagement strategies, while the least distinctive pairs—membership structures, strategic planning, and research collaboration—reflect widely shared priorities.

From a policy perspective, scientific academies must strike a balance between standardization and differentiation in their digital strategies. Aligning with recognized taxonomies ensures clarity, institutional credibility, and interoperability, while incorporating innovative elements enhances visibility and engagement. Academies with fragmented digital structures may benefit from reassessing their web organization to improve accessibility and communication effectiveness. Strengthening public outreach, ensuring transparent governance, and supporting digital transformation initiatives—particularly for academies in regions with limited resources—can help bridge disparities in digital infrastructure. Establishing international guidelines for structuring academic web content would further enhance cohesion among global academies, fostering stronger collaboration and knowledge exchange. By refining their digital presence, scientific academies can reinforce their institutional roles, expand their public reach, and strengthen their contributions to global scientific discourse.

Conclusion

This study provides a comprehensive framework for analyzing the digital presence of global scientific academies, examining how they structure their online content and engage with stakeholders. By applying Formal Concept Analysis (FCA) and graph pruning, the research identifies both common patterns and variations in the web content taxonomy of scientific academies. The findings reveal that while academies share a foundational structure emphasizing governance, knowledge dissemination, and public engagement, they differ in the extent to which they innovate or conform to established digital frameworks. The comparative analysis of taxonomy similarity, imitation scores, and innovation scores highlights distinct strategic approaches, with some academies adhering closely to conventional taxonomies, others demonstrating fragmented or underdeveloped digital strategies, and a subset actively incorporating novel structures to enhance their digital identity. The distinctiveness analysis of hypernym-hyponym pairs further provides insights into the key areas where academies differentiate themselves, reflecting diverse institutional priorities.

This study contributes to both institutional theory and digital taxonomy research. The application of FCA advances the understanding of how scientific academies navigate the balance between standardization and differentiation in their digital strategies, shedding light on institutional isomorphism in the digital realm. Additionally, the structured web mining approach and hierarchical taxonomy construction refined methods for analyzing large-scale institutional web data, offering a scalable framework for comparative analysis. These insights have practical implications for academy leaders, policymakers, and digital strategists, providing a foundation for

developing best practices that enhance the visibility, accessibility, and interoperability of scientific academies' digital presence.

Despite its contributions, this study has certain limitations. The analysis is based solely on digital content, without accounting for offline activities and interactions that may influence an academy's broader institutional role. Additionally, while taxonomy captures structural and thematic variations, it does not measure the effectiveness of digital engagement strategies. Future research could explore the relationship between digital presence and institutional influence could further refine strategies for strengthening scientific communication and global collaboration. By continuing to refine digital strategy frameworks, this research lays the groundwork for future transformations in how scientific academies facilitate scholarly communication and contribute to the global scientific ecosystem.

References

Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity? *PloS One*, *16*(4), e0249583.

Benade, L. (2016). Learned Societies, Practitioners and their 'Professional'Societies: Grounds for developing closer links. In *Educational Philosophy and Theory* (Vol. 48, Issue 14, pp. 1395–1400). Taylor & Francis.

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2024). Scraping innovativeness from corporate websites: Empirical evidence on Italian manufacturing SMEs. *Technological Forecasting and Social Change*, *207*, 123597.

Burford, S. (2011). Web information architecture–a very inclusive practice. *Journal of Information Architecture*, *3*(1), 19–40.

Burford, S. (2014). A grounded theory of the practice of web information architecture in large organizations. *Journal of the Association for Information Science and Technology*, *65*(10), 2017–2034.

Campos, P. M. C., Reginato, C. C., & Almeida, J. P. A. (2019). Towards a Core Ontology for Scientific Research Activities. In G. Guizzardi, F. Gailly, & R. Suzana Pitangueira Maciel (Eds.), *Advances in Conceptual Modeling* (Vol. 11787, pp. 3–12). Springer International Publishing. https://doi.org/10.1007/978-3-030-34146-6_1

Ceci, M., & Lanotte, P. F. (2021). Closed sequential pattern mining for sitemap generation. *World Wide Web*, *24*(1), 175–203.

Chen, X.. (2024, November). *Global scientific academies Dataset* (Version V1). Science Data Bank. https://doi.org/10.57760/sciencedb.14674

Cox, A. M. (2007). Beyond information–factors in participation in networks of practice: A case study of web management in UK higher education. *Journal of Documentation*, *63*(5), 765–787.

Cox, A. M. (2008). An exploration of concepts of community through a case study of UK university web production. *Journal of Information Science*, *34*(3), 327–345.

Elsayed, A. M. (2017). Web content strategy in higher education institutions: The case of King Abdulaziz University. *Information Development*, *33*(5), 479–494. https://doi.org/10.1177/0266666916671387

Engelbrecht, J., Djurovic, M., & Reuter, T. (2020). Current tasks of academies and academia. *Cadmus*, *4*(2), 118–126.

Engelbrecht, J., & Šlaus, I. (2022). ACADEMIES OF SCIENCES IN THE CONTEMPORARY WORLD. *Trames: A Journal of the Humanities and Social Sciences*, *26*(2), 131–139.

Gloria, M. J. K., McGuinness, D. L., Luciano, J. S., & Zhang, Q. (2013). Exploration in web science: Instruments for web observatories. *Proceedings of the 22nd International Conference on World Wide Web*, 1325–1328.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*, 653–671.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK webspace: Fifteen years of British universities on the web. *Proceedings of the 2014 ACM Conference on Web Science*, 62–70.

Isavand, L., & Poormoghim, H. (2024). Comparative Study of Scientific Academies between European Countries (Royal Society of Great Britain, Lincean Academy of Italy, French Scientific Academy), and Iran. *Advances in Applied Sociology*, *14*(03), 161–174. https://doi.org/10.4236/aasoci.2024.143011

Karanasios, S., Thakker, D., Lau, L., Allen, D., Dimitrova, V., & Norman, A. (2013). Making sense of digital traces: An activity theory driven ontological approach. *Journal of the American Society for Information Science and Technology*, *64*(12), 2452–2467.

Kenekayoro, P., Buckley, K., & Thelwall, M. (2014). Automatic classification of academic web page types. *Scientometrics*, *101*, 1015–1026.

Kenekayoro, P., Buckley, K., & Thelwall, M. (2015). Clustering research group website homepages. *Scientometrics*, *102*, 2023–2039.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLOS ONE*, *16*(4), e0249071. https://doi.org/10.1371/journal.pone.0249071

Krishnapuram, R., Joshi, A., Nasraoui, O., & Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for Web mining. *IEEE Transactions on Fuzzy Systems*, *9*(4), 595–607. https://doi.org/10.1109/91.940971

Late, E., Guns, R., Pölönen, J., Stojanovski, J., Urbanc, M., & Ochsner, M. (2024). Beyond borders: Examining the role of national learned societies in the social sciences and humanities. *Learned Publishing*.

Lepori, B., Aguillo, I. F., & Seeber, M. (2014). Size of web domains and interlinking behavior of higher education institutions in Europe. *Scientometrics*, *100*, 497–518.

Markus Neumann, Fridolin Linder and Bruce Desmarais. (2022). Government websites as data: A methodological pipeline with application to the websites of municipalities in the United States. *Journal of Information Technology & Politics*, *19*(4), 411–422. https://doi.org/10.1080/19331681.2021.1999880

Martínez-Torres, M. R., Toral, S. L., Palacios, B., & Barrero, F. (2012). An evolutionary factor analysis computation for mining website structures. *Expert Systems with Applications*, *39*(14), 11623–11633. https://doi.org/10.1016/j.eswa.2012.04.011

Norrby, E. (2001). The Role of Academies of Science in a Global World. *AMBIO: A Journal of the Human Environment*, *30*(2), 71–71.

Ruzza, M., Tiozzo, B., Mantovani, C., D'Este, F., & Ravarotto, L. (2017). Designing the information architecture of a complex website: A strategy based on news content and faceted classification. *International Journal of Information Management*, *37*(3), 166–176.

Schroeder, R., Brügger, N., & Cowls, J. (2020). Historical web as a tool for analyzing social change. *Second International Handbook of Internet Research*, 489–504.

Singh, U., Divya Venkatesh, J., Muraleedharan, A., Saluja, K. S., J H, A., & Biswas, P. (2024). Accessibility Analysis of Educational Websites Using WCAG 2.0. *Digital Government: Research and Practice*, *5*(3), 1–28. https://doi.org/10.1145/3696318

Sophia Alim. (2021). Web Accessibility of the Top Research-Intensive Universities in the UK. *Sage Open*, *11*(4), 21582440211056614. https://doi.org/10.1177/21582440211056614

Sun, A., & Lim, E. (2006). Web unit-based mining of homepage relationships. *Journal of the American Society for Information Science and Technology*, *57*(3), 394–407. https://doi.org/10.1002/asi.20279

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, *57*(1), 60–68.

Tsakalidis, A., Basile, P., Bazzi, M., Cucuringu, M., & McGillivray, B. (2021). DUKweb, diachronic word representations from the UK Web Archive corpus. *Scientific Data*, *8*(1), 269. https://doi.org/10.1038/s41597-021-01047-x

Weber, M. S. (2021). Digital Data and a Multilevel Perspective of Institutions on the Web. *Proceedings of the 13th ACM Web Science Conference 2021*, 4–4.

Will, E. M., & Callison, C. (2006). Web presence of universities: Is higher education sending the right message online? *Public Relations Review*, *32*(2), 180–183. https://doi.org/10.1016/j.pubrev.2006.02.014

Yoshinaga, N., & Nobuhara, H. (2010). Formal concept analysis based web pages classification/visualization and their application to information retrieval. *2010 10th International Symposium on Communications and Information Technologies*, 153–157.

Appendices

Table 1 presents the detailed partitioning results of the hierarchical clustering of scientific academies. Cluster 1 consists of academies that prioritize innovation, introducing novel structures and diverse content categories. This cluster includes prestigious academies from G7 and other developed countries that lead in digital strategy. Cluster 2 represents academies with fragmented or less structured digital strategies, often characterized by selective content representation or weak adherence to taxonomy standards (e.g., casciences.org, mta.hu, bas.co.bw). Cluster 3 includes academies that closely follow established taxonomies, exhibiting high imitation scores and minimal structural divergence (e.g., japan-acad.go.jp, kvab.be, vast.gov.vn).

Table 1. Detail Partition Result of the Hierarchical Clustering of Scientific Academies.

Site domain	Cluster	Site domain	Cluster	Site domain	Cluster
aast.dz	1	aciencias.org.bo	2	naskr.kg	3
ria.ie	1	ais-sanmarino.org	2	dknvs.no	3
lza.lv	1	bas.co.bw	2	igd-sh.lu	3
manu.edu.mk	1	casciences.org	2	internet.hn	3
nas.gov.ua	1	mta.hu	2	japan-acad.go.jp	3
nasb.gov.by	1	zaas.org.zm	2	knasciences.or.ke	3
nasonline.org	1	zas.ac.zw	2	kvab.be	3
nast.gov.np	1			maas.edu.mm	3
oeaw.ac.at	1			nas.go.kr	3
palast.ps	1			nas.org.ng	3
pan.pl	1			rss.jo	3
paspk.org	1			nassl.org	3
rae.es	1			nast.ph	3
ras.ru	1			nauka-nanrk.kz	3
royalacademy.dk	1			rac.gov.kh	3
lincei.it	1			sav.sk	3
royalsociety.go.th	1			sci.am	3
royalsociety.org	1			science.gov.tm	3
royalsociety.org.nz	1			snas.org.sg	3
rsc-src.ca	1			unas.org.ug	3
sanu.ac.rs	1			vast.gov.vn	3
sazu.si	1			assaf.co.za	3
science.gov.az	1			avcr.cz	3
science.org.au	1			anc.cr	3
science.org.ge	1			asrt.sci.eg	3

scnat.ch	1		acfiman.org	3
taas-online.or.tz	1		abc.org.br	3
acad-ciencias.pt	1		ac.mn	3
lma.lt	1		acaciencias.org.gt	3
tuba.gov.tr	1		academiaciencias.cu	3
leopoldina.org	1		academiadeciencias.cl	3
asm.md	1		academiadecienciasrd.org	3
kva.se	1		academie-sciences.bj	3
acad.ro	1		ashak.org	3
academy.ac.il	1		academyofcyprus.cy	3
academy.uz	1		acadsci.fi	3
academyofathens.gr	1		academie.hassan2.sciences.ma	3
akad.gov.al	1		aipi.or.id	3
akadeemia.ee	1		ansts.sn	3
akademisains.gov.my	1		asduliban.org	3
anc-argentina.org.ar	1		akademia-malagasy.mg	3
anrt.tj	1		aosci.org	3
antat.ru	1		asa.gov.af	3
anubih.ba	1		ansal.bf	3
academie-sciences.fr	1		ancperu.org	3
bas.bg	1		anciu.org.uy	3
gaas-gh.org	1		amc.edu.mx	3
knaw.nl	1		cienciasdenicaragua.org	3
bas.org.bd	1			
ias.ac.ir	1			
hazu.hr	1			
insaindia.res.in	1			
eas-et.org	1			
casinapioiv.va	1			
casad.cas.cn	1			
canu.me	1			
beitalhikma.tn	1			

The heatmap in Figure 1 visually depicts the extent to which each academy covers these core content categories. This distribution highlights clear differences in digital content strategies among academies. Some institutions, particularly those in Cluster 1, exhibit comprehensive coverage across multiple categories, whereas others, especially in Cluster 2 and Cluster 3, show gaps in specific areas, such as Public Outreach and Scientific Cooperation. The clustering approach effectively groups websites with similar digital strategies, revealing distinct content structuring behaviors across institutions.

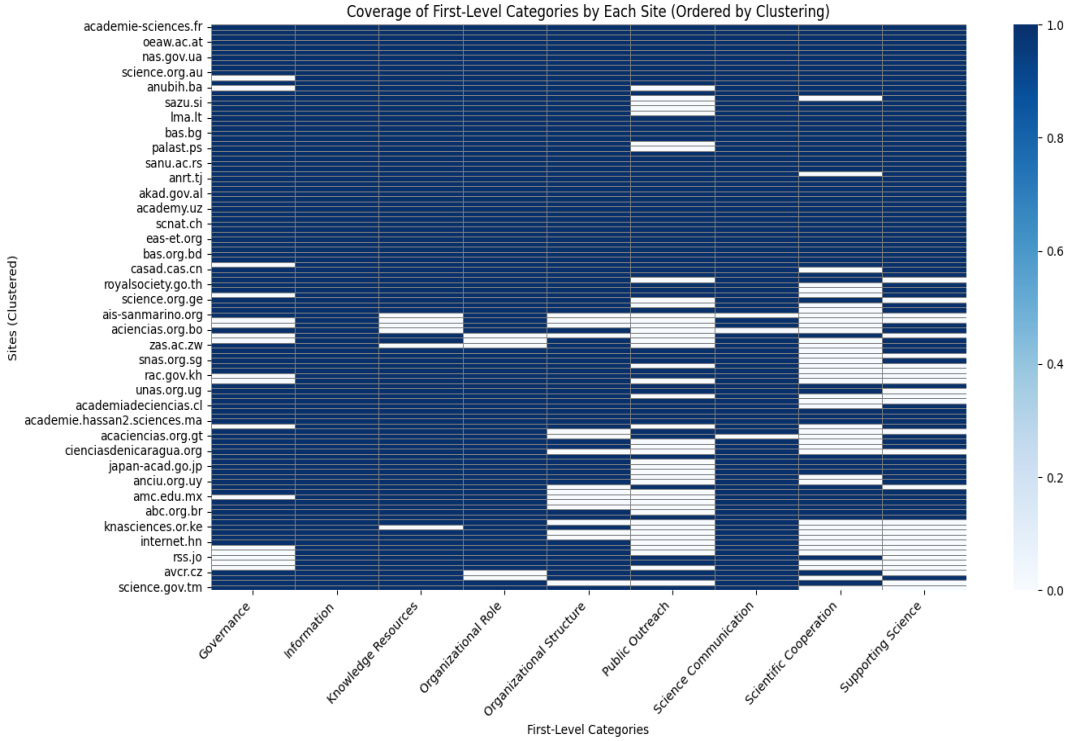


Figure 1. Comparison of First-Level Category Coverage Across Scientific Academies.

https://doi.org/10.51408/issi2025_082

What Type of Methodological Novelty is More Disruptive? Evidence from Citation Classics

Linlei Xie[1], Yi Zhao[2], Chengzhi Zhang[3]

[1] xielinlei@njust.edu.cn, [2] yizhao93@njust.edu.cn, [3]zhangcz@njust.edu.cn
Nanjing University of Science and Technology, No. 200, Xiaolingwei, 210094 Nanjing (China)

Abstract

The novel contributions of academic papers encompass various aspects such as methods, theories, and results, among which methodological novelty has been proven to be more disruptive compared to other types. Methodological novelty can be further subdivided into different types. However, which type of methodological novelty is more disruptive remains to be explored. Drawing on large language models (LLMs), this study first classifies methodological novelty in academic papers into three types: first-proposed, improvement, and application. Then, the study explores the relationship between the types of methodological novelty and disruption of scientific articles. Using 928 methodological novelty articles from Citation Classics as evidence, this study finds that first-proposed methods tend to be more disruptive, while improvement and application types tend to be less disruptive. Additionally, the study explores the effect of the number of authors and institutions on disruptiveness, finding that smaller and multi-institutional teams enhance the disruption of articles. This study explores a refined classification system for methodological novelty, aiming to enrich existing approaches to scientific innovation research and deepen understanding of novelty mechanisms.

Introduction

Measuring the novelty of papers is one of the hot topics in academic research. Novelty mainly emphasizes the difference between the research contributions in the paper and previous work, requiring that the contributions have not appeared in previous papers (Dirk, 1999). Currently, most research is limited to a quantitative measurement framework based on combination novelty theory (Uzzi et al., 2013; Wang et al., 2017). However, authors' new ideas do not always stem from atypical combinations of existing ideas (Tahamtan & Bornmann, 2018). Completely new ideas often have no discernible precedents, and fundamental breakthroughs often stem from the exploration of unknown knowledge spaces (Ahuja & Morris, 2001). At present, research on novelty mainly focuses on the novelty level of papers, with less research on novelty types. Exploring the types of novelty is particularly important and necessary, as it helps us decompose, evaluate, and measure novelty, thereby helping us better understand what novelty is and what drives it (Yan et al., 2020). The measurement of novelty degree can only capture a single dimension of it. In addition, existing articles on novel types often involve theoretical research and lack empirical exploration.

Moreover, the concepts of novelty and influence have long dominated theoretical research on scientific change, attempting to explain how new ideas change the course of knowledge (Leahey et al., 2023). Researchers have long observed that papers containing more novel ideas are more likely to be in the top 1% of citation distributions (Lee et al., 2015). Furthermore, when these novel elements are combined with an appropriate amount of conventional content, these papers are more

likely to become highly cited "hot papers" (Uzzi et al., 2013). Kuhn (1962) mentioned in his book *The Structure of Scientific Revolutions* that new ideas promote paradigm shifts in science, where a new way gradually replaces an old one. So, how do these novel ideas interact with previous work to influence future knowledge flows? Leahey et al. (2023) took a new step in this field by abandoning traditional quantitative measurement methods and dividing the novel contribution of papers into three types: new theory, new method, and new result, and deeply exploring the relationship between these types and the nature of scientific impact (measured by the CD index (Consolidating/Disruptive index, CD index) (Funk & Owen-Smith, 2017)). Leahey et al. (2023) argue that the citation count of an article can only capture the quantity of scientific impact, while the level of disruption (measured by the CD index) can better capture the nature of scientific impact, that is, the changes the article makes to the subsequent knowledge flow. Their research found that new methods tend to be more disruptive, whereas new theories tend to be less disruptive, and new results do not have a robust effect on disruptiveness; (Leahey et al., 2023). In addition, among the 2540 articles in its novelty classification dataset, there are 1459 papers on methodological novelty, accounting for over 57%.

Despite this, Leahey et al.'s (2023) typology mainly focuses on the structural level of novelty in papers and cannot distinguish specific novel ways. Evaluating how new papers can change the subsequent knowledge flow is undoubtedly a topic worth exploring in depth (Leahey et al., 2023), which can provide new insights for scientific innovation. In addition, the strong disruptiveness and dominant proportion in the classification results demonstrated by the methodologically novelty papers have also aroused our interest in further exploration. Methodological novelty can not only change the direction of knowledge flow (Leahey et al., 2023) but also the direction of scientific practice, and is often an independent foundation for future scientific discoveries (Leahey, 2008; Shi et al., 2015). Furthermore, according to the connection between methods and existing methods, methodological novelty papers can be further divided into different subtypes. Papers that propose completely new methods may be more disruptive, while papers that innovatively improve or apply existing methods may have relatively lower disruptiveness. However, the relationship between methodological novelty types and disruptiveness remains unverified in existing research. Is the high disruptiveness of methodological novelty papers caused by original methods? And what is the disruptiveness in method improvement and application-oriented articles? These questions remain to be further explored. To this end, this study will further classify methodologically novelty articles and explore the relationship between their novelty types and the essence of their scientific impact. This study aims to combine Large Language Models (LLMs) for this novel classification task.

This paper mainly studies the following two questions：

RQ1*:* How effective are LLMs in the task of classifying methodological novelty in papers？

RQ2*:* What is the relationship between different types of methodological novelty papers and disruptiveness？

Related Work

This study mainly focuses on the measurement of novelty (especially novelty classification) and its relationship with disruptiveness. So, we will review previous work from three aspects: the measurement of novelty in papers, classification, and its relationship with scientific impact.

Measurement of Novelty in Papers

For measuring the novelty of papers, researchers often develop indicators based on the logic of element novelty and recombination novelty to measure whether a paper is novel or to what extent it is novel (Kaplan & Vakili, 2015). These are mainly divided into two approaches: external indicators and internal indicators.

The measurement of novelty based on external indicators basically adopts the idea of recombinant novelty. Recombination is widely considered a source of novelty in the literature. Literature related to creativity suggests that connecting distant elements is a pathway to creativity (Uzzi et al., 2013). Management-related literature shows that a new invention stems from the synthesis of multiple ideas (Fleming, 2001; File, 2001). For academic papers, if they contain new or rare combinations of knowledge elements, they are considered novel. The main source of combinatorial novelty is the combination of previously unconsolidated elements or the combination of established elements with new concepts (Mukherjee et al., 2016). The most widely used method is to treat cited journals as a knowledge element (Uzzi et al., 2013; Tahamtan & Bornmann, 2018; Shibayama et al., 2021). If a paper cites literature from two journals that are rarely cited together, it is considered novel. Citations imply that the knowledge in the cited literature is utilized by the citing literature (Matsumoto et al., 2020). Therefore, a paper that cites a rare journal pair implies the integration of rare knowledge. Uzzi et al. (2013) proposed a method to capture the combinatorial process of research papers by calculating the relative commonality of journal pairs cited by the paper. The lowest tenth percentile commonality score in a series of commonality scores of the paper is used to measure the novelty of the paper, and the median commonality score is used to measure its conventionality. This strategy has been applied and adapted in a series of subsequent related work due to its completeness and originality. Lee et al. (2015), based on Uzzi et al.'s previous work, treated the novelty of academic papers as the scarcity of pairwise combinations of previous work (i.e., references), and measured the novelty of academic papers based on the scores of cited reference pairs. Wang et al. (2017) treated scientific research as a combinatorial process, where novelty is the exploration process of combining new knowledge with existing knowledge, and measured the novelty of science based on whether the paper is the first to combine reference journals. However, journals as a knowledge element are highly aggregated units, and citation indicators designed accordingly, although to some extent reflecting the novelty of papers (Shibayama et al., 2021), their effectiveness is still controversial (Matsumoto et al., 2020).

On the other hand, content-based novelty mainly follows two logics: one is based on element novelty, and the other is based on recombinant novelty. The main limitation of recombinant novelty is that it ignores the novelty of the knowledge elements

themselves. Completely new ideas often have no discernible precedents, and fundamental breakthroughs often stem from the exploration of unknown knowledge spaces (Ahuja & Morris Lampert, 2001). Such isolated novelty events may not be captured by recombinant novelty measures. Based on the logic of element novelty, Azoulay et al. (2001) calculated the average age of MeSH keywords to assess the novelty of articles. Some studies also believe that novelty includes both the creative development of knowledge and the inheritance and reconstruction of existing or conventional knowledge. Mishra and Torvik (2016), while exploring the relationship between MeSH terms and paper novelty, proposed that in biomedicine, a single subject term is difficult to express novelty, while combined subject terms can better reflect the novelty of papers. The most influential papers often introduce some novel combinations (atypical combinations) on the basis of traditional combinations (typical combinations). Foster et al. (2015) used entity combinations to construct a chemical knowledge network, defining the combination of knowledge entities in different clusters in the knowledge network as novel. These measurement methods are relatively intuitive, but inevitably suffer from the problem of ambiguity in textual information (such as synonyms). Although they can be solved through controlled vocabulary dictionaries, building a dictionary requires a lot of expert effort, and existing dictionaries are often domain specific.

Classification of Novelty in Papers

Regarding the classification of novelty, current classifications of novelty in papers are mainly based on two ideas. One is based on the structure and content of the article, dividing novelty types according to novel content, and the other is based on the level of novelty, dividing novelty types according to the degree of it.

Classifying articles according to novel content can be seen as a multidimensional conceptualization of novelty (Rosenkopf & McGrath, 2011), allowing us to process it more richly. Early researchers mostly based their classification standards on expert experience or questionnaire and interview results, directly classifying articles into novelty types. Dirk (1999), starting from the structure of papers, believed that if the three elements of scientific work (hypothesis, method, and result) have not been reported in previous work, scientific originality can be divided into eight types (P-P-P, P-P-N, P-N-P, N-P-P, N-N-P, N-P-N, N-N-N), and asked authors to classify their papers through questionnaires. Guetzkow et al. (2004), through interviews with panel members of scholarship competitions in social sciences and humanities, divided originality into seven types: original strategy, under-researched field, original topic, original theory, original method, original data, and original result, and found that on different dimensions of originality, both social sciences and humanities generally value the originality of methods. In addition, humanists also emphasize the originality of the data used, while social scientists appreciate more types of originality (Guetzkow et al., 2004). Heinze et al. (2009) divided originality into five types: proposing new ideas, discovering new phenomena, developing new methods, inventing new tools, and integrating existing theories from new perspectives, and invited more than 400 authoritative researchers in human genetics and nanotechnology to judge the types of 20 highly creative research results in the field.

Recently, Leahey et al. (2023) divided the novel contribution of articles into new theory, new method, and new result through rule matching, and explored which type of novelty is most disruptive to knowledge flow. The research results show that new methods are often disruptive, new theories are less disruptive, and new results have no significant relationship with scientific impact.

In addition, some researchers classify articles according to the degree of novelty. Arnqvist et al. (2013) divided the novelty of articles into high incremental, low incremental, and completely novel according to the degree of connection with existing research. Sánchez et al. (2019) divided the novelty of articles into four levels according to the degree of knowledge increment: fundamental, high incremental, incremental, and low incremental. However, whether from the perspective of novel content or degree of novelty, current classifications of novelty remain at a coarse-grained level and cannot reveal the specific ways and reasons for the novelty of articles. On the basis of Leahey et al.'s (2023) classification, we further classify methodologically novelty articles to clarify how these articles, which have strong destructive effects on subsequent knowledge flows, change the knowledge process.

Relationship Between Novelty Types and Scientific Impact

The scientific impact in this study mainly comes from peer evaluations of research and academic publications (Van, 2000). Currently, the evaluation of the scientific impact of papers is mainly through external indicators, i.e., the citation situation of the articles. However, relying solely on citation counts can only capture the quantity of scientific impact, not its nature. Therefore, some studies have begun to use citation patterns to better evaluate scientific impact (Leahey et al., 2023). It is not surprising that novel contributions often have a disruptive impact on the scientific literature. Lin et al. (2022), in a large-scale study of more than 87 million scientific papers, found that novel articles are more disruptive, with the probability of disrupting science being almost twice that of traditional papers, but this is a slow process that takes ten years or more to achieve. Ruan et al. (2023), using nearly 900,000 PubMed articles published between 1970 and 2009, measured the relationship between topic combination novelty and scientific impact, and found that topic combination novelty has an inverted U-shaped relationship with citation counts, but is positively correlated with disruptiveness. So, do different types of novel articles differ in disruptiveness? Leahey et al. (2023) have conducted related research and found that there is indeed an interesting relationship between the novelty types of articles and disruptiveness.

According to the research results of Leahey et al. (2023), methodological novelty articles are more disruptive. The portability (Porter, 1996) and wide applicability of some quantitative techniques (Abbott, 2004) promote their dissemination. New methods are often introduced from other disciplines or sub-disciplines (Abbott, 2004) and applied to problems and data related to the problem at hand. The interdisciplinary nature of most methods makes potential users unfamiliar with their foundations, and scholars who introduce and adapt methods from other fields are less constrained by existing usage conventions, so they can apply them in qualitatively new ways, resulting in more disruptive research (Leahey et al., 2023). Methods can be easily

transferred to a new environment and applied to new problems without being changed in the process (Leahey, 2005).

New theory articles are more consolidating. The new theory "requires... significant changes in conventional scientific problems and technology" (Kuhn, 1962), Therefore, they should only be constructed when existing theories can no longer explain unexpected (and cumulative) observations. In addition, "any scientific theory must be evaluated together with its auxiliary hypotheses, initial conditions, etc., especially with its predecessor, so that we can see what kind of changes it has produced" (Lakatos & Musgrave, 1970). For those who apply new theories, it is also difficult to completely separate them from their foundations (Leahey et al., 2023). A new result, even if it is truly unexpected and contradicts previous research on the topic, is unlikely to be cited alone by subsequent authors (Leahey et al., 2023). In addition, new results (usually generated at the active research "frontier") must be linked to existing theories (residing in the consensus and paradigm "core") to be recognized and understood (Cole, 1983).

We already know that articles that are novel in different elements have different disruptiveness, but in fact, even within a single type of novel article, there are still different novel patterns and strategies. For example, for methodological novelty articles, some propose an unprecedented method, some improve existing methods, and some articles innovatively apply existing methods. So, for different novel patterns within the same type of novel article, will there be some differences in disruptiveness? Subjectively, papers proposing completely novel methods should be more disruptive than improvement and application-type papers, but this speculation has not been verified. At present, no researchers have explored the potential relationship, so this paper intends to further distinguish different novel patterns in methodological novelty articles, to deeply explore how such articles change subsequent knowledge flows.

Data and Methodology

Since this study is based on Leahey et al.'s (2023) research to further subdivide the novelty types of methodological novelty papers and explores the relationship between their different subtypes and disruptiveness, we adopt the same method as theirs to first divide papers into three types: new theory, new method, and new result, and obtain the methodological novelty articles required for this study. Specifically, we use Citation Classics essays as the data source, use the synonym dictionary developed by Leahey et al. (2023), and adopt a rule-based method to obtain three types of novel sentences, thereby performing article-level novelty classification. For disruptiveness, we use the CD index (Consolidating/Disruptive index, CD index) developed by Funk and Owen-Smith (2017) and employed by Leahey et al. (2023) to measure. Leahey et al. (2023) also mentioned in the article that subsequent scholars have used this measure and re-labeled it the “disruption index” (Bornmann et al. 2020; Wu et al. 2019), but it is equivalent to Funk and Owen-Smith’s (2017) CD index. Moreover, the CD index has demonstrated robust performance across multiple validation tests conducted by Funk and Owen-Smith (2017), as well as subsequent studies adopting the metric, such as research by Wu et al. (2019) and

Azoulay et al. (2020). There are actually other indicators for measuring disruptive behavior. For example, Chen (2006) proposed Freeman's Betweenness centrality, but this indicator is suitable for identifying key nodes of cross domain connections and revealing the mediating role of knowledge flow. There are also FV index (Prabhakaran et al., 2015) and FV gradient (Lathabai et al., 2015; Prabhakaran et al. 2018), which rely on complex network path analysis and have high computational costs. In addition, there are also the multidimensional evaluation framework proposed by Bu et al. (2021) and the semantic based evaluation method proposed by Yan and Fan (2024), which lack the simplicity and practicality of the CD index and cannot effectively measure changes in knowledge flow. The CD index uses citation patterns to quantify the degree to which a focal paper increases or decreases its dependence on its predecessor papers (i.e., its cited references). The logic of its calculation is that papers with a stronger consolidating impact should increase their citations to predecessor papers, while papers with a greater disruptive impact should do the opposite. Since this indicator quantifies whether and how a paper changes the knowledge flow on which it is based, it can be conceptualized as a scientific impact indicator (Leahey et al., 2023). In this way, this study will deeply explore how methodological novelty papers with high disruptiveness and a large proportion of articles change subsequent knowledge flows. The research framework is shown in Figure 1:

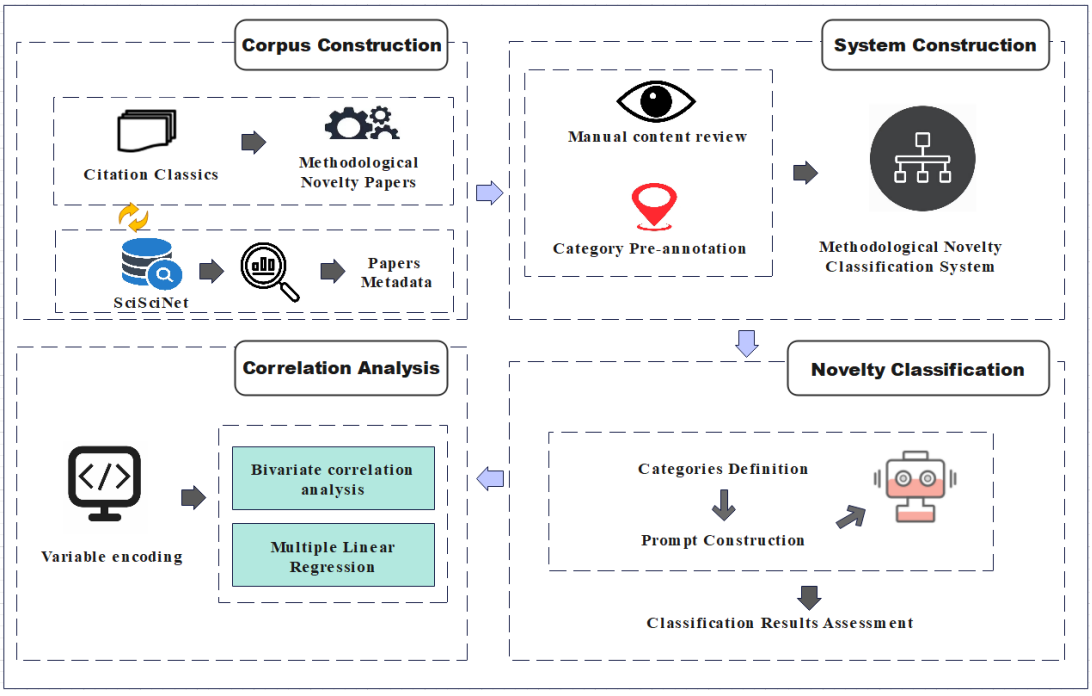


Figure 1. Framework of this study.

Data

This study uses Citation Classics as the data source. Citation Classics refer to journal articles published between 1936 and 1987 that have been cited more than a specified

number of times in Web of Science[1]. Citation Classics essays are written by Citation Classics authors and are solicited by Eugene Garfield, the developer of Web of Science. Many years after the original papers were published, these authors were invited to write a short (about one page) essay reviewing the origins of their projects, the challenges they encountered, and the reasons they believe their work has had a profound impact. These Citation Classics essays encourage scientists to "construct their own contributions," promoting the production of "intellectual self-narratives" (Gross, 2008), so they contain rich sociological information. These essays provide us with a rare humanized perspective on science, which is rarely seen in traditional journal articles or bibliometric metadata. These one-page essays span 17 years (1977 to 1993) and cover all major scientific fields. Leahey et al. (2023) OCR scanned the Citation Classics essays to form text files. By constructing a synonym table, they used a rule-based method to identify different types of novel sentences in Citation Classics essays and aggregated them at the article level to obtain the novelty type of each article. We mainly conducted further novelty classification of methodological novelty articles, we use the same method as Leahey et al. (2023) to obtain the novelty classification dataset, and separated 1459 methodologically novelty articles from it for further novelty classification in this study.

It is worth mentioning that to test whether the views of Citation Classics authors are consistent with those of the scientific community, Leahey et al. (2023) obtained the "citation context" of each Citation Classics article studied from the Microsoft Academic Graph (MAG)[2] to understand how other papers expressed themselves when citing these classic articles. They used regression models and confusion matrices to compare the Citation Classics author perspective (collected from Citation Classics essays) and the scientific community perspective (collected from MAG citation contexts), and the results confirmed that the two are consistent in their views on the novelty types of articles. This to some extent also ensures the reliability of the methodological novelty data used for further classification in this study.

Due to the limited access to Web of Science resources, we were unable to obtain the metadata of some articles, so we decided to link our data with the SciSciNet database and matched a total of 1226 articles. SciSciNet is a large-scale open data lake for the science of science research, covering over 134M scientific publications and millions of external linkages to funding and public uses (Lin et al., 2023). In addition, we further obtained the corresponding metadata of the articles through the DOI, including the title, journal name, publication year, author names and affiliations, number of co-authors and institutions, and the CD index used to measure disruptiveness.

Classification System for Methodological Novelty in Papers

This study mainly further distinguishes different novel patterns in methodological novelty articles to obtain subtypes of methodological novelty, and on this basis, explores their relationship with disruptiveness. Therefore, on the basis of Leahey et

[1] https://www.webofscience.com/wos

[2] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph

al.'s (2023) data, we extract methodological novelty articles to construct the corpus to be classified in this study.

There is relatively little research on the further classification of methodological novelty articles. German scientist Mensch divided novel contribution into three types according to importance: Basic novelty, Improving novelty, and Fake novelty (Mensch, 1979). Among them, Basic novelty marks the beginning of previously unknown new products or new processes based on new scientific principles; Improving novelty refers to minor but important improvements to products, processes, and services; and Fake novelty refers to external modifications to products or processes that do not lead to changes in their consumer characteristics. In addition, the National Science Board (US) divided novelty into incremental and transformative according to the way science develops (2011). Arnqvist et al. (2013) also divided the novelty of articles into high incremental, low incremental, and completely novel according to the degree of connection with existing research. This paper synthesizes previous classifications of novelty and combines the characteristics of academic paper research methods to divide the types of methodological novelty in papers into three subtypes: First-proposed, Improvement, and Application. The category definitions are shown in Table 1.

After formulating the classification system for methodological novelty papers in this study, we manually annotated 100 articles for the construction of the subsequent evaluation dataset and as examples to be added to the Prompt to help LLMs better understand the classification task.

Table 1. Definitions of Methodological Novelty Types in Papers.

Novelty Type	*Definition*	*Example*
First-proposed	This method is first proposed in this paper and has never appeared in other scientific works before. This method is not an improvement or application of other methods, nor is it a combination of several other methods.	This paper described the first completely automatic method for colorimetric analysis.
Improvement	This method is an improvement or modification of methods that have appeared in previous scientific works, or a combination of previously proposed methods.	Our first attempts to improve the method used the incredibly laborious ion chamber technique.
Application	This method is the introduction or application of methods already proposed in previous scientific works.	It was the first attempt to apply one of the numerical hydrodynamic codes to the problem of the collapse and explosion of a star.

Classification Method for Methodological Novelty in Papers

In recent years, LLMs have shown significant progress in various challenging tasks, including solving mathematical problems (Romera-Paredes et al., 2024), proving mathematical theories (Wang et al., 2023a), and generating code to solve analytical or computational tasks (Huang et al., 2024). These advances have opened up new possibilities for using LLMs to accelerate research (Wang et al., 2023b), including research on novelty classification (Huang et al., 2025). For the novelty classification task in this study, we adopt the method of using LLMs with Few-Shot Prompting. This study uses deepseek_v3[1], llama-3[2], qianwen-2[3], and gpt-3.5-turbo[4] models to further classify methodological novelty papers. By comparing the classification results of multiple models, the model with the best performance was selected to participate in the subsequent regression analysis with the CD index.

Specifically, referring to Huang et al.'s (2025) research on LLMs, we design a prompt to elucidate the criteria and methodology for novelty classification of methodological novelty articles, and added some examples to the Prompt to assist the LLMs in understanding the methodological novelty classification task. Before using LLMs for formal classification, we randomly selected some articles and conducted LLMs classification and manual category labeling, and manually reviewed and compared the results. By analyzing the erroneous data identified by LLMs, we iteratively improve the Prompt to make it clearer in terms of task and novelty category definition, thereby improving the performance of LLMs in this task. Table 2 shows the specific content of the final Prompt, which mainly includes three parts: "###Instruction", "###Input", and "###Output". LLMs classify novelty based on the relevant text of the input article and provide classification criteria by understanding the definition of methodological novelty types and learning from a small number of annotated examples.

[1] https://www.deepseek.com
[2] https://www.llama.com
[3] https://tongyi.aliyun.com
[4] https://openai.com

Table 2. Prompt for Methodological Novelty Classification.

###Instruction	As a proficient scholar, your task is to evaluate the methodological novelty of a given paper based on the definitions of methodological novelty and its types, as well as analyzing the provided artificial classification examples. **Definition of Methodological Novelty** Methodological novelty refers to the extent to which a scientific output contributes to the knowledge of a particular research field in terms of methods. Methodological novelty exists in anything that adds new things to the knowledge of the method in the field. Methodological novelty can be first proposed, improved on existing methods, application-oriented, or even a mixture of them. **Definition of Methodological Novelty Type:** **1.First-proposed:** this method was first proposed in this paper and has never appeared in other scientific works before. This method is not an improvement or application of other methods, nor is it a combination of several other methods. **2.Improvement:** this method is an improvement or modification of methods that have appeared in previous scientific works, or a combination of previously proposed methods. **3.Application:** this method is the introduction or application of methods already proposed in previous scientific works. Please provide Methodological Novelty Type and Methodological Novelty Description. **Methodological Novelty Type:** choose Methodological Novelty Type of the given paper from [First-proposed, Improvement, Application]. If there are no suitable options, output 'None'. **Methodological Novelty Description:** write a concise paragraph (no >500 words) to explain the reasons for choosing Methodological Novelty Type. **Examples of classification of Methodological Novelty :** **Example 1:** **Sentence:** This paper described the first completely automatic method for colorimetric analysis. **Methodological Novelty Type:** First-proposed **Example 2:** **Sentence:** Our first attempts to improve the method used the incredibly laborious ion chamber technique. **Methodological Novelty Type:** Improvement **Example 3:** **Sentence:** It was the first attempt to apply one of the numerical hydrodynamic codes to the problem of the collapse and explosion of a star. **Methodological Novelty Type:** Application
###Input	**the relevant text of a given paper:** …
###Output	**Methodological Novelty Type (MNT):** … **Methodological Novelty Description (MND):** …

"###Instruction" helps LLMs understand the conceptual basis, analysis methods, and goals of the novelty evaluation of methodological novelty articles. In terms of conceptual basis, a concise definition of methodological novelty is proposed: "Methodological novelty refers to the extent to which a scientific output contributes

to the knowledge of a particular research field in terms of methods. Methodological novelty exists in anything that adds new things to the knowledge of the method in the field. Methodological novelty can be first proposed, improved on existing methods, application-oriented, or even a mixture of them." According to this definition, Methodological Novelty Types (MNT) include three categories: First-proposed, Improvement, and Application, and detailed definitions of these three categories are given. In this part, three manually classified examples are added to help LLMs understand the classification task. To achieve the evaluation goal, the LLM needs to provide the Methodological Novelty Type (MNT) and Methodological Novelty Description (MND).

"###Input" includes the relevant text of the given paper. This text is the corresponding Citation Classics essay, i.e., the one-page author's self-narrative content, which is regarded as the author's self-construction of the article's contribution many years later.

"###Output" includes MNT and MND, which are generated by the LLMs. MNT represents the novelty type of the methodological novelty paper, which can be first-proposed, improved, or application-oriented; MND is a concise paragraph to clarify the reasons for assigning the MNT.

Correlation Analysis Techniques Between Methodological Novelty Types and Disruptiveness

First, we set all three novelty types as binary variables. If an article is classified into that type, the variable is coded as 1, otherwise as 0. To deeply explore the potential relationship between novelty types in methodological novelty papers and disruptiveness, we first conducted an independent samples t-test on the relationship between method novelty types and CD index to determine whether there is a significant difference in the mean CD index between two sample populations belonging to and not belonging to a certain methodological novelty type.

To control for the influence of other variables such as the size of the paper team and the mutual influence between our three independent variables, we further conducted a multiple linear regression with the three methodological novelty types as independent variables and the CD index as the dependent variable. We built a multi-level linear regression model by successively adding control variables.

Since some articles are quite old, some articles in the SciSciNet database have publication years beyond the original time range of Citation Classics (1931-1987), and there may also be some errors in the publication year of articles in the database. To ensure the accuracy of our results, we decided not to include the publication year of articles as a control variable in our regression analysis model. Specifically, this study included key control variables to ensure the accuracy and reliability of the analysis results. The control variables include:

- Article type: whether the article is a journal paper, conference paper, or other.
- Number of authors: referring to the research of Singh & Fleming (2010) and Wu et al. (2019), the number of authors is used as an indicator of team

collaboration, which may affect the diversity and depth of novelty achievements.

- Number of institutions: institutions, as support for resources and technical elements, may affect the advancement and research depth of methodological novelty achievements.

Results

Due to the fact that there are 1226 articles in our data that can be found in the SciSciNet database for corresponding article records when obtaining other metadata of the article, of which 928 articles have a CD index. Therefore, we further extracted these articles to participate in the subsequent regression analysis.

Evaluation of the Classification Results of Methodological Novelty in Papers

We answer RQ1 in this section. We first reviewed the fine-grained classification results of methodological novelty of the four models: deepseek_v3, llama-3, qianwen-2, and gpt-3.5-turbo. We randomly selected 100 articles and manually annotated the methodological novelty categories to construct an evaluation dataset to evaluate the classification results. The evaluation mainly includes two parts: one is the accuracy of the Methodological Novelty Type (MNT) classification, and the other is the completeness and logic of the Methodological Novelty Description (MND) content. This study selected Precision(P), Recall(R), and F_1 score as evaluation indicators for MNT. These indicators are calculated based on the "confusion matrix", which is 2 * 2 for the most common binary classification. It includes True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), as shown in the table 3.

Table 3. confusion matrix.

Real / *Predict*	*1*	*0*
1	TP	FP
0	FN	TN

The detailed definitions of the above indicators are as follows:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$\mathrm{F1} = 2 * \frac{Precision * Recall}{Precision+Recall} \quad (3)$$

The evaluation comparison of the MNT classification results of the four models is shown in Table 4.

Table 4. Evaluation of Methodological Novelty Classification Results.

Model	*P*	*R*	F_1
deepseek_v3	0. 718	0. 710	0. 705
llama-3	**0. 815**	**0. 810**	**0. 804**
qianwen-2	0. 723	0. 670	0. 662
gpt-3.5-turbo	0. 554	0. 540	0. 513

Through comparison, we found that llama-3 outperforms the other three models in the accuracy of MNT. We further evaluated MND, taking the classification result of one article as an example, as shown in Table 5.

Table 5. Example of LMS Model Classification Results.

ocr_article	*deepseek-v3*	*llama-3*	*qianwen-2*	*gpt-3.5-turbo*
Number 51 Citation Classics December 19, Singer J M & Plotz C M. The latex fixation test. I. Application to the serologic diagnosis of rheumatoid arthritis. American Journal of Medicine 21:888-92, 1956. The authors describe a latex fixation test for the serologic diagnosis of rheu- matoid arthritis which improves on standard tests depending on erythrocyte agglutination by using biologically in - ert polyvinyl toluene and polysterene latex particles of uniform size.… (Due to space limitations, the remaining parts of this article have been omitted. For details, please refer to: https://garfield.library.upenn.edu/classics1977/A1977EC98100002.pdf)	###MNT: **Improvement** ###MND: The paper describes the development of the latex fixation test, which improves upon existing methods for the serologic diagnosis of rheumatoid arthritis. The authors replaced the previously used erythrocyte agglutination method with biologically inert polyvinyl toluene and polystyrene latex particles of uniform size. This modification addressed the limitations of earlier methods, such as the nonspecific reactions observed with collodion particles. The novelty lies in the use of latex particles, which were first developed in 1954, and their application to create a more reliable and standardized diagnostic test. While the latex particles themselves were not novel, their specific application in this context, combined with the identification of the antigen as human IgG, represents a significant improvement over prior techniques. Thus, the method novelty type is classified as ""Improvement"" because it builds upon and refines existing methodologies rather than introducing an entirely new concept."	###MNT: **Improvement** ###MND: The paper describes the development of a latex fixation test for the serologic diagnosis of rheumatoid arthritis, which improves on standard tests that depend on erythrocyte agglutination. The authors utilized biologically inert polyvinyl toluene and polystyrene latex particles of uniform size, which was a novel approach at the time. This improvement is built upon previous work, such as the use of collodion particles and the identification of the antigen as human IgG. The introduction of latex particles of uniform size, developed by Backus and Vanderhoff, was a key factor in the development of this test. The combination of this new latex with previous findings led to the creation of the latex fixation test, which has remained a standard in the field. The improvement lies in the use of a more reliable and efficient method for serologic diagnosis, making it a significant contribution to the field of immunology. Note: Although the paper builds upon previous work and introduces a new method, it is not considered ""First-propose	###MNT: **Improvement** ###MND: The method described in this paper represents an improvement on existing methods rather than a first-proposed or purely application-oriented approach. The authors, Morgan and Lazarow, developed a two-step procedure for radioimmunoassay of insulin by building upon prior research. Specifically, they utilized the foundational work of Moloney & Coval (1955), who demonstrated that guinea pigs could be reliably immunized with mammalian insulins, and Arquilla & Stavitsky (1956), who showed the feasibility of using immunological procedures for insulin assays. Additionally, Skom & Talmage (1958) had reported on the use of anti-human gamma globulin to precipitate non-precipitating insulin antibodies in insulin-resistant human serum. By integrating these previous findings, Morgan and Lazarow introduced a novel two-antibody system: in the first step, insulin forms a soluble complex with its specific antibody from immunized guinea pigs; in the second step, this complex is precipitated by an antibody to guine	###MNT: **Improvement** ###MND: The method described in the paper is an improvement on standard tests for the serologic diagnosis of rheumatoid arthritis. By utilizing biologically inert polyvinyl toluene and polystyrene latex particles of uniform size, the authors improved on the standard methods that depended on erythrocyte agglutination. This method represents an enhancement or modification of existing methods, making it a clear example of an Improvement in terms of Method Novelty."

	d"" because it improves upon existing methods r ather than proposing a c ompletely new concept. It is also not solely an "" Application"" because it involves significant novel ty and improvement in t he method itself, rather t han just applying an exis ting method to a new pr oblem. Therefore, the m ost appropriate classifica tion is "Improvement".	a pig serum obtained fro m immunized rabbits. Th is method significantly i mproved the precision a nd reproducibility of insu lin measurement, especi ally for samples containi ng less than a nanogram of insulin. The novelty li es in the combination an d refinement of existing techniques to create a m ore efficient and reliable immunoassay method. Therefore, the method n ovelty type is classified a s "Improvement" becaus e it builds on and enhan ces previously establishe d methodologies rather than proposing a compl etely new or unexplored method.

The article "describes a latex fixation test for the serologic diagnosis of rheumatoid arthritis, which improves on standard tests depending on erythrocyte agglutination by using biologically inert polyvinyl toluene and polystyrene latex particles of uniform size." Our several models all correctly classified this article as the 'Improvement' type.

For MND, the llama-3 model gave the result: "This paper describes the development of a latex fixation test for the serologic diagnosis of rheumatoid arthritis, which improves on standard tests that depend on erythrocyte agglutination. The authors utilized biologically inert polyvinyl toluene and polystyrene latex particles of uniform size, which was a new approach at the time. This improvement is built upon previous work, such as the use of collodion particles and the identification of the antigen as human IgG. The introduction of latex particles of uniform size, developed by Backus and Vanderhoff, was a key factor in the development of this test. The combination of latex with previous findings led to the creation of the latex fixation test, which has remained a standard in the field. The improvements are in the use of a more reliable and efficient method for serologic diagnosis, making it a significant contribution to the field of immunology. Note: Although the paper builds upon previous work and introduces a new method, it is not considered 'First-proposed' because it improves upon existing methods rather than proposing a completely new concept. It is also not solely an 'Application' because it involves significant novelty and improvement in the method itself, rather than just applying an existing method to a new problem. Therefore, the most appropriate classification is 'Improvement'."

It can be seen that llama-3 clearly and completely gave the reasons for classifying the article as 'Improvement', as well as the reasons for not classifying it as the other two novelty types, and the narrative is also more logical. Llama-3 not only gave the specific improvements and novel contributions of the article but also provided relevant information about the original method. In contrast, we found that other models only gave reasons for not classifying it as 'First-proposed' and did not give

reasons for not classifying it as 'Application', and their narratives were not as logical as llama-3.

In addition, although we only limited the Prompt to no more than 500 words, by comparing the output results of the models, we found that the gpt-3.5-turbo model's classification results were much shorter than other models, with an average of only about 80 words; deepseek_v3 had an average of about 150 words; llama-3 model and qianwen-2 model had an average of about 250 words.

Therefore, considering the evaluation results of MNT and MND, we finally selected the classification results of the llama-3 model as the final methodological novelty classification results of this study.

Descriptive Statistical Results of Methodological Novelty Types

According to the classification results of llama3, among the 928 articles, 191 were classified as First-proposed, 572 were classified as Improvement, and 146 were classified as Application; 19 articles were judged by the model as MNT being None, meaning they were not in our three categories. Table 6 shows the descriptive statistical results of our data, including the control variables involved in this study.

Table 6. Descriptive Statistical Results of Methodological Novelty Classification.

Variable \ *Metric*	*Mean*	*SD*	*Min.*	*Max.*
CD Index	0.15	0.23	-0.16	0.99
MNT				
First-proposed	0.21	0.41	0	1
Improvement	0.62	0.49	0	1
Application	0.16	0.36	0	1
Controls				
Team_Size	2.41	1.69	1	16
Institution_Count	1.17	0.59	1	9
Doc_Type(1- Journal;0-other)	0.97	0.17	0	1

We found that there are more disruptive papers in our data, with less consolidating papers. This may be because the Citation Classics we selected are often highly cited and influential. However, as this article aims to explore how innovative methods can change the way subsequent research cites focused papers, that is, to investigate the impact of three types of methodological novelty on disruptiveness. So this is not a problem in our research. The overall disruptiveness in the articles is not very high, with an average of only 0.15. However, there are significant differences in disruptiveness between articles, with the CD index of the article with the highest disruptiveness reaching 0.99.

In our classification results, the proportion of 'Improvement' articles is the highest, reaching 62%; Next is 'First-proposed', accounting for 21%, with the lowest proportion being 'Application' type. In addition, the number of collaborating scholars in different studies varies greatly, with the largest team size reaching 16 people, and

the standard deviation reaching 1.69; but the mean is 2.41, which means that the collaboration team usually consists of 2-3 people. The number of collaborating institutions corresponding to different articles also varies, but the number is generally small, usually consisting of 1-2 institutions.

Correlation Analysis Results Between Methodological Novelty Types and Disruptiveness

We answer RQ2 in this section. We first conducted an independent samples t-test, and the results are shown in Table 7.

Table 7. Independent Samples t-test Results Between Methodological Novelty Types and CD Index.

	Yes	*No*
First-proposed	0.199*** (n = 191)	0.135 (n = 737)
Improvement	0.147 (n = 572)	0.150 (n = 356)
Application	0.092 (n = 146)	0.159*** (n = 782)

Note: *p < .05;**p < .01;***p < .001.

According to the t-test results, articles belonging to the 'First premise' category (meanCD =0.199) are more disruptive than articles not belonging to this category (meanCD =0.135), and are significant at the p=.001 level; There is no significant difference in disruptiveness between articles that belong to and do not belong to the category of 'Improvement'; On the contrary to 'First premise', articles that do not belong to 'Application' (meanCD =0.159) are more disruptive than articles that belong to this type (meanCD =0.092), and are significant at the p=.001 level.

Due to the previous t-test not considering control variables, more accurate results need to be further estimated using multiple linear regression with the addition of control variables. In Model 1, we only studied the correlation between the three methodological novelty types and the CD index; Model 2 added "Number of Collaborating authors" (Team_Size); Model 3 added the control variable "Number of Collaborating Institutions" (Institution_Count); Model 4 added all control variables. Table 8 shows the results of the multiple linear regression analysis.

Table 8. Multiple Linear Regression Results of Methodological Novelty Types and CD Index.

	Model 1	Model 2	Model 3	Model 4
MNT				
First-proposed	0.130*	0.128*	0.276**	0.276**
	(0.055)	(0.054)	(0.099)	(0.099)
Improvement	0.077	0.078	0.239*	0.241*
	(0.053)	(0.053)	(0.097)	(0.097)
Application	0.026	0.028	0.162	0.165
	(0.055)	(0.055)	(0.098)	(0.098)
Controls				
Team_Size		-0.012**	-0.022***	-0.023***
		(0.005)	(0.006)	(0.006)
Institution_Count			0.065***	0.064***
			(0.019)	(0.019)
Doc_Type (1-Journal;0-other)				0.109*
				(0.055)
R^2	0.020	0.027	0.054	0.061
N	928	928	928	928

Note: *p < .05;**p < .01;***p < .001. Standard errors in parentheses.

According to the results of Model 1, when only considering the relationship between the three methodological novelty types and the CD index, we found that all three methodological novelty types are positively correlated with the CD index, but only the relationship between 'First-proposed' and the CD index is significant (b = 0.130*); the coefficient of 'Improvement' is slightly smaller than that of 'First-proposed' (b = 0.077), and the coefficient of 'Application' is even smaller (b = 0.026). The newly proposed method has no basis in the original method, so when subsequent scholars cite this article, many will not choose to cite the references of this article as supplementary discussions of the method, which leads to the generally higher CD index of this type of article. For 'Improvement' and 'Application' type methodological novelty articles, authors often cite related articles of the original method when introducing them, and subsequent scholars will also cite the references of this article when citing it to better introduce the principle of the method or to clarify the founder of the method, which leads to their CD index being relatively smaller than that of 'First-proposed' type articles. Therefore, although the first-proposed method has risks due to its uncertainty, once successful, its return is often very high, which can change subsequent knowledge flows and significantly promote the development of the field.

When we added the control variable 'Team_Size' in Model 2, the results still hold, 'First-proposed' is still positively correlated with the CD index and significant (b = 0.128*); the positive correlation between the other two novelty types and the CD index is still not significant, and the coefficients are smaller than that of 'First-proposed'; moreover, we found that the number of co-authors is negatively correlated with the CD index and statistically significant (b = -.012**). This result is also consistent with the conclusion of Wu et al. (2019), whose research is based on large-scale papers, patents, and software products data with various levels of influence. We have also obtained consistent conclusions in high impact Citation Classics datasets. To test whether this negative correlation is due to an inverted U-shaped correlation between the number of co-authors and the CD index, we squared the value of 'Team_Size' and participated in the regression analysis with the CD index, but the results showed that there is no inverted U-shaped correlation between the squared number of co-authors and the CD index.

We further added a control variable 'Institution_Cunt' in Model 3. We found that the positive correlation between 'First-proposed' and the CD index still holds, and the significance level has increased (b = 0.276**); in addition, we surprisingly found that the positive correlation between 'Improvement' and the CD index becomes significant (b = 0.239*), and the number of co-authors is still negatively correlated with the CD index and significant (b = -0.022**). Moreover, the number of collaborating institutions is positively correlated with the CD index and particularly significant (b = 0.065***). The more resources and broader research networks brought by multi-institutional collaboration may be the reason for this relationship.

We ultimately added the control variable 'Doc_Type' in Model 4. It can be seen that the positive correlation between the 'First proposed' type and the CD index still holds and is relatively significant (b=0.276**); the positive correlation between 'Improvement' and CD index still holds and is significant (b=0.241*); The relationship between 'Application' and CD index is still not significant. The negative correlation between 'Team_Size' and the CD index, as well as the positive correlation between 'Institution_Cunt' and the CD index, still hold and are significant. In addition, we found that journal articles are more disruptive than non journal articles. Similar to the results of Leahey et al. (2023), the impact of our methodological novelty type on disruptiveness is statistically significant, but also small. However, in reality, it is difficult to explain highly complex results such as the CD index, which rely on citation behavior not only by the authors of the article, but also by the broader scientific community (Leahey et al., 2023). Leahey et al. (2023) also converted the CD index into percentile of disruptiveness as Wu et al. (2019) did and found that the impact was comparable in scale to the team size coefficient they proposed. Moreover, our main focus is on the comparison between the three types of methodological novelty. Overall, the most compelling conclusion we have drawn is that there exists a significant positive correlation between the "First-proposed" type and the CD index, and disruptiveness of this type of articles is significantly higher than the other two types. Furthermore, disruptiveness of "Improvement" type is also higher compared to "Application" type. This result is consistent with our cognition and hypothesis.

Discussion

In this study, the high disruptiveness shown by papers proposing new methods is consistent with the cognition of the scientific community. Research proposing methods for the first time has no original method as a basis and may completely disrupt existing research paradigms or introduce completely new concepts. These methods often break existing knowledge frameworks and have greater potential to promote changes in scientific knowledge flows. On the other hand, improvement and application-type research is more about optimizing and expanding on existing knowledge, with less impact on subsequent knowledge flows.

In addition, smaller teams have advantages in both communication costs and decision-making processes, allowing them to adjust research directions more quickly, thereby helping to produce more disruptive research results. Large teams may be more inclined to adopt more conservative research methods to reduce risks and ensure the stability and reproducibility of research. The complexity of multi-scholar collaboration may hinder the implementation of innovative ideas in the research process, thereby reducing the disruptiveness of research.

Multi-institutional collaboration can integrate more resources, such as experimental equipment, data sets, and funding, and the integration of these resources helps to carry out more complex and innovative research. In addition, multi-institutional collaboration usually involves a broader research network, and different institutions may focus on different research fields. This makes multi-institutional collaboration more likely to come into contact with more research frontiers and emerging fields, and combine the latest advances in different fields to produce disruptive research results.

Theoretical Implications

Overall, this study has the following theoretical implications.

Firstly, this study enriches the research content of novelty evaluation in articles. This study is the first to propose dividing methodological novelty articles into three subtypes: first-proposed, improvement, and application. This classification method not only enriches the research content of articles' novelty evaluation but also provides a more detailed analysis framework for subsequent research. Through this classification, researchers can more deeply understand the unique characteristics and impacts of different types of novel articles. By deeply exploring the underlying mechanisms of articles' novelty, this study improves the interpretability of methodological novelty. This mechanism analysis helps to reveal the internal logic of novelty generation and provides a theoretical basis for future research.

Secondly, it expands the research perspective of scientometrics. This study analyzes the relationship between the novelty types of methodological novelty articles and their disruptiveness (CD index), thereby exploring which type of methodological novelty can better change subsequent knowledge flows. This research not only expands the research horizon of scientometrics but also provides a new perspective for understanding the dissemination and evolution of scientific knowledge. By exploring the impact mechanisms of different types of novel articles on knowledge flows, this study provides a new research direction for the field of scientometrics and

helps to further understand the role of scientific novelty in promoting the knowledge system.

Furthermore, this study validates the feasibility of artificial intelligence technology in classifying articles' novel contribution. This study applied advanced artificial intelligence technology, especially LLMs, to the task of article novelty classification, verifying its feasibility in handling complex text classification tasks. This application not only expands the methods of articles' novelty classification but also provides references for other text analysis tasks. The generalization ability and complex feature capture ability of LLMs make up for the shortcomings of manual classification caused by personal disciplinary background and subjective factors. This technical application improves the accuracy and scientificity of classification results and provides reliable tools for future research.

Practical Implications

The practical implications of this study can be summarized in the following three aspects.

Firstly, optimize scientific research management and policy-making. By understanding the impact and disruptiveness index of different types of novel articles (first-proposed, improvement, application), scientific research managers and policy makers can more scientifically allocate resources, prioritize support for research with high disruptiveness and potential impact, thereby maximizing the return on scientific research investment. In addition, the results of this study can provide a basis for the formulation of scientific research policies and educational training programs, encourage cross institutional and interdisciplinary cooperation, and support high-risk and high return research projects. Moreover, in talent cultivation, special attention should be paid to original thinking and abilities, heuristic teaching should be encouraged, and innovative practical activities should be carried out in a timely manner.

Secondly, improve the academic evaluation system. Traditional academic evaluation systems usually rely on quantitative indicators such as citation counts. Although they can reflect the dissemination scope and influence of research, they are difficult to accurately measure the novelty of research. By introducing the novelty classification of methodological novelty articles, the academic evaluation system can be improved, and the novelty, influence, and long-term value of research results can be more comprehensively evaluated. The positive correlation between the novelty type and disruptiveness of methodological novelty articles derived from research can also motivate researchers to engage in more innovative and disruptive research, encourage exploration of unknown fields, and promote scientific progress.

Thirdly, promote scientific research cooperation and achievement transformation. This study found that cross-institutional collaboration helps to produce more disruptive research results. Therefore, scientific research managers and policy makers can promote more cross-institutional and cross-disciplinary collaborative projects, promote knowledge sharing and resource integration. By identifying scientific research achievements with high disruptiveness potential, scientific research institutions can accelerate their transformation and application, promote the

combination of scientific and technological novelty and economic development, and deepen industry-university-research collaboration.

Limitations

The focus of this study is to explore the further classification of methodological novelty articles. To ensure the quality of methodological novelty articles, our research is based on the first-level classification of novelty by Leahey et al. (2023), so the data scale is relatively small. In addition, since methodological novelty articles have strong disruptiveness, we have only further divided the novelty types of such articles at present, and the further classification of theoretical novelty and result novelty remains to be explored. Furthermore, our methodological novelty classification only utilizes the currently popular and widely recognized four LLMs models with good performance, and adopts a Few-Shot Prompting approach. The performance of other classification methods and models in this classification task still needs further exploration. In addition, the novelty classification method of this study largely depends on the clear statements made by the authors of Citation Classics when reviewing the paper. But authors may implicitly describe their methodological contributions without explicitly labeling them as "first-proposed" "improvement" or "application, or may use outdated or different terminology. This situation may not be well captured and correctly classified by large models. Finally, this article used retrospective essays from Citation Classics (often decades-old papers). Since these texts are reflections written many years after original publication, the original authors' descriptions and terminology choices may no longer align clearly with present-day understandings. Terms and concepts once considered novel or groundbreaking can become standard practice or even obsolete over time. When contemporary LLMs interpret these historical reflections, they probably do so with the knowledge patterns learned from more recent textual corpora, potentially misclassifying or overlooking nuances related to past methodological innovations. This is a potential limitation of our research.

Conclusion and Future Work

This study is the first to divide methodological novelty into three types: first-proposed, improvement, and application, and introduces LLMs to classify methodological novelty. Through independent samples t-test and multiple linear regression analysis, the impact of different types of methodological novelty on disruptiveness is revealed. The study found that articles proposing new methods for the first time have higher disruptiveness, while improvement and application-type articles have relatively lower disruptiveness. In addition, we found that the number of co-authors has a significant negative correlation with disruptiveness, while the number of collaborating institutions has a significant positive correlation with disruptiveness.

At present, this study has only further classified methodological novelty papers, and will subsequently explore other novelty categories, such as the novelty classification of theoretical novelty. In the future, we will also conduct our novelty classification experiments on a larger scale of data to verify the universality of the results of this

study. And combined with more complex machine learning models to improve the accuracy and efficiency of articles' novelty classification. In the future, we will continue to study how to better explore the potential "novelty descriptions" in papers using LLMs, thereby improving the performance of LLMs in novelty classification. In addition, we noticed that methodological novelty articles are often more disruptive, which may be closely related to their portability (Porter, 1996) and interdisciplinary nature (Abbott, 2004). Therefore, articles in different disciplines may have significant differences in the way they change subsequent knowledge flows. Therefore, in the future, we will also combine disciplinary differences for more in-depth exploration.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.72074113) and Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX24_0792). Special thanks to Dr. Jina Lee for sharing the relevant data and code.

References

Abbott, A. D. (2004). Methods of discovery: Heuristics for the social sciences.

Ahuja, G., & Morris Lampert, C. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic management journal*, 22(6-7), 521-543.

Arnqvist, G. (2013). Editorial rejects? Novelty, schnovelty! *Trends in ecology & evolution*, 28(8), 448-449.

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, 42(3), 527-554.

Azoulay, P., Jones, B. F., Kim, J. D., & Miranda, J. (2020). Age and high-growth entrepreneurship. American Economic Review: Insights, 2(1), 65-82.

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020). Disruptive papers published in Scientometrics: Meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). Scientometrics, 123(2), 1149-1155.

Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. Quantitative science studies, 2(1), 155-183.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology, 57(3), 359-377.

Cole, S. (1983). The hierarchy of the sciences? *American Journal of sociology*, 89(1), 111-139.

Dirk, L. (1999). A measure of originality: The elements of science. *Social Studies of Science*, 29(5), 765-776.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management science*, 47(1), 117-132.

File, D. (2001). The nber patent citation data file: lessons, insights and methodological tools.. *NBER Working Paper*, *8498*, 40.
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American sociological review*, 80(5), 875-908.
Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management science*, 63(3), 791-817.
Gross, N. (2008). *Richard Rorty: The making of an American philosopher*. University of Chicago Press.
Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is Originality in the Humanities and the Social Sciences? *American Sociological Review*, 69(2), 190-212.
Heinze, T., Shapira, P., Rogers, J. D., & Senker, J. M. (2009). Organizational and institutional influences on creativity in scientific research. *Research Policy*, 38(4), 610-623.
Huang, Q., Vora, J., Liang, P., & Leskovec, J. (2024). MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. *International Conference on Machine Learning*.
Huang, S., Huang, Y., Liu, Y., Luo, Z., & Lu, W. (2025). Are large language models qualified reviewers in originality evaluation? *Information Processing & Management*, 62(3), 103973.
Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36(10), 1435-1457.
Kogabayev, T., & Maziliauskas, A. (2017). The definition and classification of innovation. *HOLISTICA–Journal of Business and Public Administration*, 8(1), 59-72.
Kuhn, T. S. (1962). *The structure of scientific revolutions* (Vol. 962). Chicago: University of Chicago press.
Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the growth of knowledge: Volume 4: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Cambridge university press.
Lathabai, H. H., Prabhakaran, T., & Changat, M. (2015). Centrality and flow vergence gradient based path analysis of scientific literature: A case study of biotechnology for engineering. Physica A: Statistical Mechanics and its Applications, 429, 157-168.
Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*, 84(1), 1-24.
Leahey, E. (2008). Methodological memes and mores: Toward a sociology of social research. *Annu. Rev. Sociol*, 34(1), 33-53.
Leahey, E., Lee, J., & Funk, R. J. (2023). What types of novelty are most disruptive? *American Sociological Review*, 88(3), 562-597.
Lee, Y. N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research policy*, 44(3), 684-697.
Lin, Y., Evans, J. A., & Wu, L. (2022). New directions in science emerge from disconnection and discord. *Journal of Informetrics*, 16(1), 101234.

Lin, Z., Yin, Y., Liu, L., & Wang, D. (2023). SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1), 315.

Matsumoto, K., Shibayama, S., Kang, B., & Igami, M. (2020). A validation study of knowledge combinatorial novelty.

Mensch, G. (1979). *Stalemate in Technology*. Cambridge Massachusetts: Ballinger Publishing Company.

Mishra, S., & Torvik, V. I. (2016). Quantifying Conceptual Novelty in the Biomedical Literature. *D-Lib magazine: the magazine of the Digital Library Forum*, 22(9-10), 10.1045/september2016-mishra. https://doi.org/10.1045/september2016-mishra.

Mukherjee, S., Uzzi, B., Jones, B., & Stringer, M. (2016). A new method for identifying recombinations of existing knowledge associated with high-impact innovation. *Journal of Product Innovation Management*, 33(2), 224-236.

National Science Board (US). (2011). *National Science Foundation's merit review criteria: review and revisions*. National Science Foundation.

Prabhakaran, T., Lathabai, H. H., & Changat, M. (2015). Detection of paradigm shifts and emerging fields using scientific network: A case study of Information Technology for Engineering. Technological Forecasting and Social Change, 91, 124-145.

Prabhakaran, T., Lathabai, H. H., George, S., & Changat, M. (2018). Towards prediction of paradigm shifts from scientific literature. Scientometrics, 117, 1611-1644.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., ... & Fawzi, A. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625(7995), 468-475.

Rosenkopf, L., & McGrath, P. (2011). Advancing the conceptualization and operationalization of novelty in organizational research. *Organization Science*, 22(5), 1297-1311.

Ruan, X., Ao, W., Lyu, D., Cheng, Y., & Li, J. (2023). Effect of the topic-combination novelty on the disruption and impact of scientific articles: Evidence from PubMed. *Journal of Information Science*, 01655515231161133.

Sánchez, I. R., Makkonen, T., & Williams, A. M. (2019). Peer review assessment of originality in tourism journals: critical perspective of key gatekeepers. *Annals of Tourism Research*, 77, 1-11.

Shi, F., Foster, J. G., & Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43, 73-85.

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PloS one*, 16(7), e0254034.

Singh, J., & Fleming, L. (2010). Lone inventors as sources of breakthroughs: Myth or reality? Management science, 56(1), 41-56.

Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of informetrics*, 12(3), 906-930.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468-472.

Van Houten, B. A., Phelps, J., Barnes, M., & Suk, W. A. (2000). Evaluating scientific impact. *Environmental health perspectives*, 108(9), A392-A393.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8), 1416-1436.

Wang, H., Xin, H., Zheng, C., Li, L., Liu, Z., Cao, Q., ... & Liang, X. (2023a). Lego-prover: Neural theorem proving with growing libraries. *arXiv preprint arXiv:2310.00656*.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., ... & Zitnik, M. (2023b). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972), 47-60.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. Nature, 566(7744), 378-382.

Yan, Y., Tian, S., & Zhang, J. (2020). The impact of a paper's new combinations and new components on its citation. *Scientometrics*, 122, 895-913.

Yan, Z., & Fan, K. (2024). An integrated indicator for evaluating scientific papers: considering academic impact and novelty. Scientometrics, 129(11), 6909-6929.

https://doi.org/10.51408/issi2025_019

Where Did Post-Doctorates Go? A Factorized Analysis on Chinese Postdoctoral Program for Innovative Talent

Tan Fu[1], Wen Lou[2]

[1]52284419013@stu.ecnu.edu.cn
Department of Information Management, Faculty of Economics and Management, East China Normal University, Shanghai (China)

[2]wlou@infor.ecnu.edu.cn
Department of Information Management, Faculty of Economics and Management, East China Normal University, Shanghai (China)
Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education of China, Shanghai (China)

Abstract

Young scientific and technological talents, as the core force of scientific research and innovation, have increasingly drawn academic attention regarding their career trajectories and the effects of policy interventions. The postdoctoral experience is becoming an indispensable stage. This study, based on empirical data from China's National Postdoctoral Program for Innovative Talent (NPPIT), systematically analyzes the basic characteristics, academic mobility, and title promotion of the NPPIT fellows by integrating scientific metrology methods and multiple logistic regression models. The findings are as follows: (a)There is a gender imbalance among NPPIT postdocs, with the age group predominantly ranging from 27 to 31 years. The majority of doctoral institutions are Project 985 universities. Already, 16.05% of the postdocs have obtained senior-level treatment, and the proportion of fellows securing tenure-track positions after program completion is higher than 65%. (b)The academic mobility exhibits significant stratification: the migration rate of postdocs from 985 universities to 211 and other general universities reaches 30.94%, reflecting the "competitive crowding-out effect" and the trend of resource reallocation. (c)The accelerated effect of the promotion path: 52.74% of early-funded fellows achieve senior titles within 6-8 years, indicating that the NPPIT significantly shortens the professional cycle. (d)Title and the type of doctoral institution are significant factors influencing academic mobility, while gender and tenure do not show significant correlations.(e)Alma mater sentiment plays a role in career choices, with many NPPIT postdocs choosing to stay at or return to their undergraduate or doctoral institutions, although it is not the decisive factor.(f)The academic mobility of NPPIT postdocs reflects the competitive academic job market and the importance of institutional reputation and resources in shaping career decisions.The contribution of this study lies in revealing the interactive effects of institutional factors, such as the pre-tenure and tenure systems and individual strategies, such as alma mater sentiment. Additionally, the study offers policy recommendations for optimizing the postdoctoral training system, including hierarchical evaluation, mobility incentives, and data-driven decision-making.

Introduction

Scientific and technological innovation serves as the core driving force of social and economic development, a process heavily reliant on the innovative capabilities of young scientific and technological talents. Classic studies indicate that scientists' productivity between the ages of 35 and 40 accounts for more than 70% of their significant lifetime achievements (Lehman, 1953). Young scientific and technological talents, defined as individuals under 40 years old who are in the early stages of their careers and possess significant research potential (Chen, 2022; Li et

al., 2024), have a direct impact on the effectiveness of national science and technology strategies (Zhang et al., 2024). Although countries universally cultivate young scholars through funding programs, such as the "Career" program by the U.S. NSF and the Starting Grants by the European Research Council (ERC), China's unique National Postdoctoral Program for Innovative Talent (NPPIT) has received little attention from the international academic community.

The program not only represents the highest national recognition of postdoctoral research capabilities but also establishes early identification criteria for young scientific and technological talents through a "selecting the best from the best" mechanism. Since its implementation in 2016, it has cumulatively funded more than 3,300 top postdoctoral researchers under the age of 31(approximately 1% of China's total postdoctoral recruits), providing a unique sample for analyzing the growth paths of young scientific and technological talents. Analyzing the characteristics of its fellows can offer empirical evidence for addressing the "35-year-old anxiety" among young talents and optimizing postdoctoral training policies. Existing scientometric research often focuses on mature talent programs such as the Nobel Prize (Rodríguez, 2022; Chan et al., 2018), NSFC Distinguished Young Scholars Fund and Excellent Young Scholars Fund (Li et al., 2024; Liu et al., 2022; Yuan et al., 2018; Yin et al., 2017), but there is a lack of systematic analysis of NPPIT fellows.

As a core force in national scientific and technological innovation, the postdoctoral community plays an irreplaceable strategic role in promoting academic progress, fostering interdisciplinary integration, and responding to global technological competition (Ma, 2023). Especially against the backdrop of the “Double First-Class ” initiative and the strategy of innovation-driven development, postdocs serve not only as the main force in university research but also as a vital bridge for international academic exchange (Liu et al., 2023). Moreover, the postdoctoral system functions both as a talent cultivation mechanism and a regulator of the academic labor market, making its dynamic evolution and optimization pathway crucial to enhancing national technological competitiveness. Therefore, studying the postdoctoral community is not only related to individual career development but also involves systematic issues of higher education governance, research innovation ecosystems, and talent policies.

Over the past decade, the global postdoctoral scale has expanded significantly, but the imbalance between supply and demand in the academic labor market has intensified, making career prospects uncertain a widespread challenge (Gao et al., 2022). In China, despite the postdoctoral experience being proven to significantly increase the probability of obtaining elite faculty positions by enhancing the quality and impact of research outputs (Xu et al., 2024), postdocs still face multiple challenges in their professional development: ambiguous role positioning, such as the conflict between “teacher” and “student” identities (Jiang et al., 2024), low job satisfaction with only 40% satisfied with the academic environment (Zhu, 2014), insufficient economic security (Yang et al., 2024), and mental health risks. Additionally, international comparisons show that Chinese postdocs exhibit a "low investment-high utilization" model, with their professional development capabilities significantly below the global average (Zhao et al., 2023), while overseas experience

significantly enhances their competitiveness in the academic market. These realities highlight the urgency of optimizing the postdoctoral system and improving the professional ecosystem.

Existing research on postdoctoral fellows predominantly employs quantitative analysis, with a minority using qualitative methods such as regression models (Liu et al., 2022), text analysis, and mixed methods. Some studies have also introduced cross-national comparisons and policy evidence-based analysis (Liu X et al., 2023) to enhance the universality and practical orientation of the conclusions. The core issues can be summarized into the following four categories: (a) Career Pathways and Market Returns: This focuses on the impact of postdoctoral experience on academic promotion, revealing the heterogeneity of postdoctoral experience on faculty position acquisition through tracking data analysis and propensity score matching (Ye et al., 2024). (b) Institutional Design and Training Effectiveness: Qualitative analysis explores the management of postdoctoral mobile stations, funding systems, and the optimization of classification and evaluation mechanisms (Chen et al., 2023; Ma, 2023). (c) Mental Health and Organizational Support: Based on structural equation modeling, this analyzes how mentor support and job meaning mitigate professional burnout, emphasizing the critical role of organizational support and psychological capital (Jiang et al., 2022, Zhao et al., 2022; Cai et al., 2022). (d) Role Conflict and Identity: Utilizing role theory and in-depth interviews, this deconstructs the tension of multiple identities of postdoctoral fellows and their institutional roots (Li et al., 2019; Song et al., 2022).

In general, the existing research tends to focus on how postdoctoral experiences influence the acquisition of academic positions, the best practices for optimizing training systems, and strategies for mental health interventions. However, there are three prominent limitations in the current body of research: Firstly, the unique growth trajectories of elite postdoctoral scholars, particularly their academic mobility and its determinants, are often overlooked. Secondly, there's a lack of exploration into the cumulative benefits that early-stage research projects, such as NPPIT, confer on postdoctoral scholars' academic careers. Thirdly, there's an excessive dependence on cross-sectional data, which hinders the thorough analysis of longitudinal career data. These research gaps present opportunities for this study to delve into. Specifically, will postdoctoral scholars funded by NPPIT secure academic positions? If not, what institutions do they move to? What factors influence their academic mobility? And what do these movements reveal about the academic landscape? This study aims to provide insightful answers to these questions.

Data and methods

Firstly, the list of NPPIT fellows for the years 2016-2024 was obtained. Starting from November 2024, our research team downloaded the list of NPPIT fellows for the years 2016-2024 from the official website of the China Postdoctoral Science Foundation, which included names, host institutions, primary disciplines, and funding numbers. Since the official website of the Postdoctoral Science Foundation no longer publicizes the names of fellows in defense and military systems from 2022 onwards, there has been no public channel to obtain the list of fellows in these sectors.

Subsequently, through various means such as personal homepages, institutional official websites, search engines, the China National Knowledge Infrastructure database, author searches in Web of Science, ORCID, ResearchGate, and others, we gathered the curriculum vitae information of 3,371 NPPIT postdocs from 2016 to 2024, including gender, date of birth, current institution, current department, PhD award date, PhD institution, field of study, work experience, master's degree award date, master's institution, bachelor's degree award date, and undergraduate institution. The information collection period was from November 2024 to January 2025. Finally, the collected NPPIT postdoc curriculum vitae (CV) information was input into a unified format data table in preparation for subsequent data processing and cleaning; the categorized information was quantitatively encoded to construct a comprehensive postdoctoral innovative talent CV database.

The postdoctoral experience is increasingly becoming a necessary condition for young innovative talents pursuing academic careers. Typically, most doctoral students start their academic careers after graduation, and in the context of a scarcity of faculty positions, a postdoctoral position is the optimal choice. It provides a transitional and cumulative opportunity for PhD holders, allowing for periods of free exploration that can lead to formal faculty positions, associate senior titles, or even senior titles. What characteristics are common among postdoctoral innovative talents who successfully secure promotions? Factors such as job changes, reasons for job choices, and educational backgrounds are crucial for observing the mobility of young innovative talents. Gender, age, educational background, field of study, and academic mobility are important factors for observing the promotion of young innovative talents. Chi-square tests and multiple logistic regression analyses were used to examine the relationship between these variables and title levels. Considering the research questions and the presence of missing data, we selected the CV information of 2,468 NPPIT postdocs with clearly defined positions as the sample for analyzing the mobility and promotion of young innovative talents.

A scientific CV is a true reflection of a researcher's academic career, documenting their growth trajectory, including fields of study, educational level, institutional changes, international experience, research output, and collaborative teamwork. It provides a new method and perspective for the study of postdoctoral innovative talent policies. Existing research indicates that gender, age, academic background, international experience, institutional nature, and frequency of mobility are important variables affecting talent development. However, since NPPIT applicants are required to be under 31 years old with similar lengths of education, most NPPIT postdocs are of similar age, so this study does not focus on age as a primary indicator but instead uses the time since funding was received as an important grouping variable for observation. Therefore, this study uses gender, educational origin, and academic mobility as the main indicators for empirical analysis of the group characteristics of NPPIT postdocs and their relationship with growth.

Results

Overall analysis

Firstly, we conducted statistics on the funding years, hosting institutions, and disciplines of the 3,371 NPPIT postdocs in 2016, there were 200 fellows, 300 in 2017, and 400 each year from 2018 to 2021. In 2022, 2023, and 2024, there were 367, 450, and 454 fellows, respectively. Among these NPPIT postdocs, 2,116 were affiliated with Project 985 institutions (China's initiative to build world-class universities, launched in 1998), accounting for 62.77%, the; 485 were affiliated with the Chinese Academy of Sciences (CAS), accounting for 14.39%; 432 were affiliated with Project 211 institutions (the national program focused on developing 100 key universities and disciplines for the 21st century, initiated in 1995), accounting for 12.82%; 162 were affiliated with other Chinese universities, accounting for 4.81%; 128 were affiliated with other Chinese research institutes, accounting for 3.80%; and 39 entered the defense and military industries, and 9 entered Chinese enterprises, accounting for 1.16% and 0.27% respectively. The top ten institutions with the most NPPIT postdocs were Tsinghua University, Peking University, Fudan University, University of Science and Technology of China, Shanghai Jiao Tong University, Zhejiang University, Xi'an Jiaotong University, Sun Yat-sen University, Wuhan University, and Tongji University. The majority of their postdoctoral experiences belonged to the natural sciences, with the top five disciplines being biology, materials science, chemistry, clinical medicine, and physics. Only 4 NPPIT postdocs belonged to the social sciences, specifically 1 in psychology, 2 in applied economics, and 1 in statistics.

Secondly, we statistically analyzed the gender, ages in funding year, current institution, and current title of the 2,468 NPPIT postdocs with clearly defined positions, as shown in Table 1 and Figure 1. In terms of gender distribution, the proportion of women in the NPPIT postdocs group was small, with only 526 females, accounting for 21.31%, which is similar to the gender ratio of recipients of China's Excellent Young Scientist Fund Program (Chen, 2022). The trend of female representation among high-level talents is consistent with the observation that higher talent levels have fewer women, with the proportion of female Excellent Young Scientists ranging from 18.32% to 23.33% between 2012 and 2020. However, as shown in Figure 1(a), the proportion of women has shown a fluctuating increase over time, indicating a positive trend in the proportion of female young innovative talents. Regarding birth year and age at the time of receiving NPPIT, the Postdocs were born between 1985 and 1998, with ages in funding year ranging from 27 to 31 and an average of 29.1 years, with exceptions such as Associate Professor Wang Panding from Beijing Institute of Technology, who was selected for NPPIT at the age of 24.

Table 1. The overall distribution of sample NPPIT postdocs (*p<0.05, **p<0.01, ***p<0.001).

Variable		*Total (N=2,468)*	*Unanalyzed subsample (N=1,104)*	*Mobility subsample (N=1,364)*		
				Mobility group (N=422)	*Non-mobility group (N=942)*	*Test*
Gender (%)	Male	78.69	76.90	77.49	81.32	χ^2=2.45, p=0.12
	Female	21.31	23.10	22.51	18.68	
Avg.age in funding year (Mean ± SD, years)		28.98 (1.49)	29.29 (1.50)	28.76 (1.48)	28.96 (1.52)	t=0.68, p=0.50
Current institution (%)	Institutes	14.71	20.20	12.80	9.13	χ^2=17.31, p=0.0006 ***
	Other univs	9.12	5.89	23.22	6.58	
	Project 211	17.38	9.33	27.73	22.19	
	Project 985	58.79	64.58	36.26	62.10	
	Postdoc	44.73	100.00	0.00	0.00	
Current title (%)	Junior	0.16	0.00	0.24	0.32	χ^2=14.72, p=0.002 **
	Intermediate	8.31	0.00	12.09	16.35	
	Deputy senior	30.75	0.00	35.78	57.32	
	Senior	16.05	0.00	51.90	26.01	
Tenure (%)		44.45	0.00	79.15	80.68	χ^2=0.34, p=0.56

Regarding the current institutions and types, the top ten institutions hosting the largest number of NPPIT postdocs are Tsinghua University, Peking University, Fudan University, Xi'an Jiaotong University, Zhejiang University, University of Science and Technology of China, Shanghai Jiao Tong University, Sun Yat-sen University, Tongji University, and Wuhan University. In this context, we categorize universities overseas and those in Hong Kong, Macau, and Taiwan as "other universities," which represent all universities outside of Project 211 and Project 985. CAS, other Chinese research institutes, Chinese enterprises, and the defense and military industries are collectively referred to as "research institutes." As shown in Figure 1(b), over time, there is a trend towards NPPIT postdocs being increasingly affiliated with more Project 985 universities, with a decrease in the proportion of Project 211 universities and other universities, and a symmetrical fluctuation in the proportion of research institutes, remaining stable at the beginning and end. However, looking at the NPPIT recipients from a reverse chronological perspective, it suggests a future trend where NPPIT recipients from 2020 to 2024 are likely to move to Project 211 universities and other universities.

Overall, among the 2,468 NPPIT postdocs 1,451 are currently affiliated with Project 985 universities, accounting for 58.79%, the highest proportion; 429 with Project 211 universities, accounting for 17.38%; 363 with research institutes, accounting for 14.71%, of which CAS accounts for 11.06%; and the other research institutes, Chinese enterprises, and the defense and military industries account for 3.16%,

0.41%, and 0.08% respectively. Additionally, 225 are affiliated with other universities, accounting for 9.12%, including 8.83% Chinese other universities, 0.20% overseas universities, and 0.08% universities in Hong Kong, Macau, and Taiwan. Compared to the proportion of postdoctoral institution types, the proportion of Project 985 universities has decreased, and if we consider only the NPPIT postdocs who completed their programs from 2016 to 2021, this proportion would drop to 54.89%. The proportion of CAS has decreased, while the proportions of Project 211 universities and other Chinese universities have increased. This is related to the employment situation for postdoctoral fellows, as there are fewer lifetime tenure positions in Chinese universities and research institutes, with many adopting the international practice of fixed-term contracts. Project 985 universities and CAS have abundant resources and better research conditions but high assessment requirements and intense competition, leading to those who fail assessments moving to Project 211 universities and other universities. Data also shows that the proportion of NPPIT postdocs moving to Hong Kong, Macau, Taiwan, and overseas is low, indicating that the postdoctoral innovation support program is effective in supporting young scientific and technological talents.

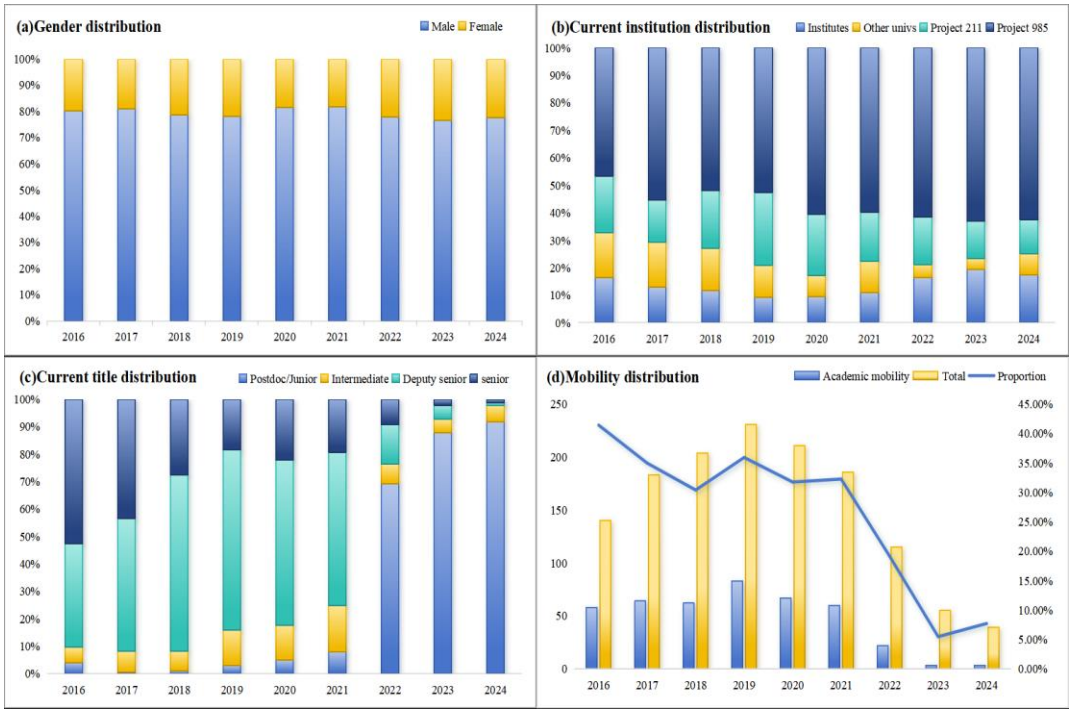


Figure 1. The overall and mobility distribution of sample NPPIT postdocs.

In terms of current titles and whether they hold lifetime tenure, this study includes both permanent and non-permanent senior and associate senior positions in the new title system. Overall, 396 NPPIT postdocs have achieved senior positions, accounting for 16.05%, while 759 have achieved associate senior positions, accounting for 30.75%. Intermediate titles account for 8.31%, and there are 1,108 at

the junior level or still in postdoctoral positions, accounting for 44.89%. Over time, as shown in Figure 1(c), the proportions of senior and associate senior positions among NPPIT postdocs from 2016 to 2021 are both higher than 75%, indicating that the titles of NPPIT postdocs who received funding earlier have significantly improved. The proportion of senior positions among the 2016 NPPIT postdocs has reached 52.74%, and comparing this with the NPPIT postdocs from 2022 to 2024, it is evident that half of the NPPIT postdocs can achieve senior positions within 6-8 years, which is much faster than the average 10-12 years typically required for postdoctoral fellows to reach full professorship (Liet al., 2017; Jensen et al., 2009), demonstrating the significant effect of NPPIT on cultivating innovative talents. It should also be noted that a certain proportion remain at the junior level or in postdoctoral positions after completing NPPIT, and there are significant differences in promotion among NPPIT postdocs of the same year. As shown in Figure 1(d), the proportion of NPPIT postdocs obtaining lifetime tenure positions after completing the program is greater than 65%, exceeding half, and this proportion is expected to increase over time, indicating that most NPPIT postdocs can obtain relatively stable positions quickly and have a rapid promotion trend.

Academic mobility

For those whose current title is still postdoctoral, it is generally because they have not changed their position, such as those undergoing a second postdoctoral term at the same institution or those who have not yet completed their project. Therefore, we will select the 1,364 NPPIT postdocs with a current title other than postdoctoral from the 2,468 NPPIT postdocs as our sample to explore the mobility and promotion of NPPIT postdocs

By comparing the names of the NPPIT hosting institutions and the current institutions of the sample, we found that among the 1,364 NPPIT postdocs from 2016 to 2024, 422 have different postdoctoral institutions from their current institutions, accounting for 30.94% of the total, indicating a significant scale of mobility. As shown in Figure 2(a), except for an increase in the number of migrants in 2019, the number of migrants generally decreases with the increase in funding year. The 2016 NPPIT postdocs have the highest proportion of migrants at 41.43%, while the mobility rate dropped to 19.13% in 2022. The academic mobility in 2023 and 2024 is an exceptional phenomenon, where postdocs entered the institution first and received funding after 1-2 years, so they had already left the station and entered another institution by the time of our study.

In terms of gender, out of 422 postdoctoral fellows, 327 were male and 95 were female. In the total mobile population, males accounted for 77.49% and females accounted for 22.51%. The male proportion was 54.98 percentage points higher than the female proportion, indicating that male academic mobility was more prevalent. However, considering the large gender disparity in the total sample, the calculation shows that the proportion of mobile female postdoctoral fellows within the female sample was 35.06%, while for males it was 29.92%, suggesting that females are more inclined towards academic mobility.

Looking at the geographical flow of mobility, only three postdoctoral fellows moved internationally: from Fudan University to the Chinese University of Hong Kong, from Beijing University of Aeronautics and Astronautics to Coventry University in the UK, and from Fudan University to the University of Texas at Austin. International mobility among postdoctoral fellows is rare. According to China's seven geographical divisions, Figure 2(a) shows that the net outflow was highest in the North and Northeast regions, while the inflow regions were more or less similar. The net inflow in the East, South, Central, and Southwest regions was around 20, indicating that the overall flow trend is from the North and Northeast to other regions. However, in terms of numbers, the North and East are the two regions with the highest postdoctoral mobility. Within these regions, outstanding young scholars tend to move short distances. Only the Northeast has a low number of mobilities and a negative net flow, with only 2 out of 13 people (15.3%) moving within the Northeast region, and 84.7% moving out of the Northeast, indicating a serious loss of outstanding postdoctoral fellows in the region. The North region, with many universities in Beijing, has the largest number of postdoctoral fellows, but due to intense competition in Beijing and less abundant university resources in other cities, there are not many cities and opportunities to accept postdoctoral fellows. Unlike the Northeast, the East region has several important cities for economic development such as Shanghai, Nanjing, and Hangzhou, so the North region experiences serious academic brain drain, but this is more like a density dispersion, sending talent all over the country. Out of the 206 people in the North region who engaged in academic mobility, 100 people (48.5%) moved within the North region, nearly half, while 41 people (19.9%) moved to the East region, 25 people (12.1%) to the Central region, 21 people (10.1%) to the South region, 8 to the Southwest region, 7 to the Northwest region, and 3 to the Northeast region. The outflow from the East region was also mainly internal, accounting for 61.7%. Moreover, statistics on the cities of mobility show that Beijing is the city with the highest number of people moving, with 200 people moving out, nearly half of the total mobility. Intra-city mobility in Beijing reached 46%, with the rest evenly distributed to major cities across the country, highlighting the contribution of Beijing's universities to the national talent supply. Shanghai is the second-highest city with 63 people moving out and 43 moving in, indicating some talent loss but not severe, and the numbers are far less than Beijing. In other cities with high mobility numbers, the inflow and outflow are balanced, suggesting that young talents serving as postdoctoral fellows in these cities tend to move within the city or nearby, maintaining a stable talent pool of postdoctoral fellows.

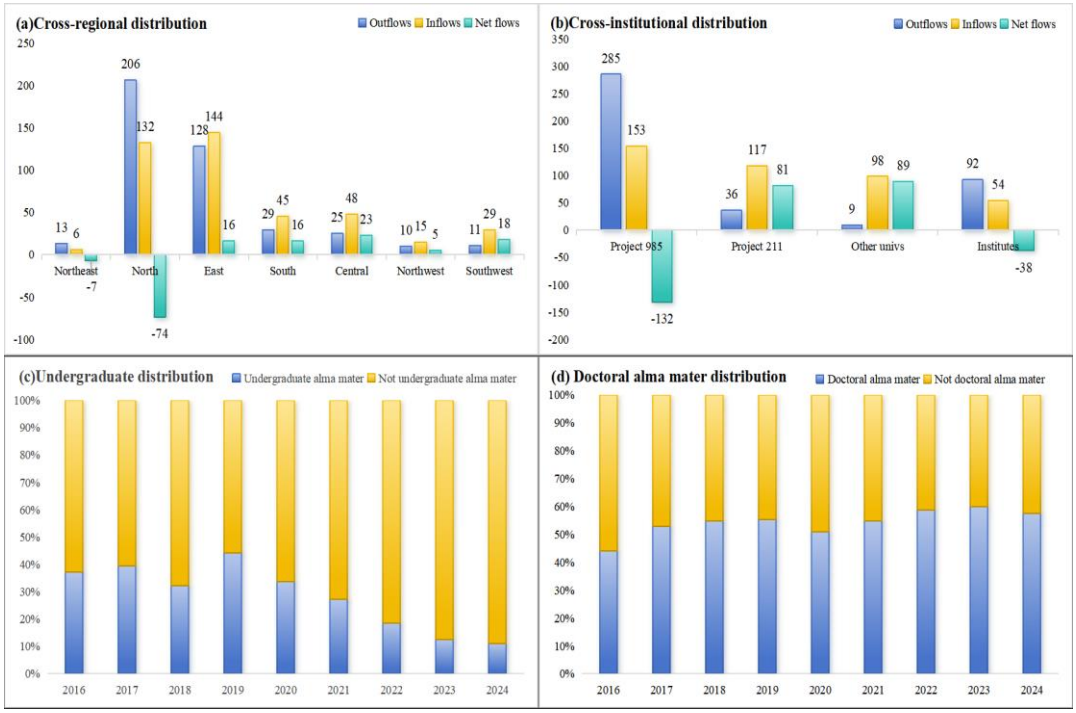


Figure 2. Academic mobility distribution of sample NPPIT postdocs.

As shown in Figure 2(b), in terms of institution types, there is a significant loss of NPPIT postdocs from Project 985 universities and research institutes, with most moving to other universities and Project 211 universities. Of the 285 NPPIT postdocs from Project 985 universities, 85.26% moved to Chinese universities, including 30.53% to Project 211, 20.35% to other Chinese universities, and 34.39% to other Project 985 universities. Additionally, 12.98% moved to research institutes, 3 to universities in Hong Kong, Macau, and overseas, and 2 to Chinese enterprises. Among them, Tucunchao, a 2018 NPPIT fellow from Tsinghua University, became the founder of Power Law Intelligence. Of the 92 NPPIT postdocs from research institutes, 47.83% moved to Project 985 universities, 21.74% to other Chinese universities, 20.65% to Project 211, 7.61% to other research institutes, and 1.09% each to the defense military and Chinese enterprises.

The top five institutions with the highest outflow of NPPIT postdocs are Peking University, Tsinghua University, Fudan University, Shanghai Jiao Tong University, and the University of Science and Technology of China. The top five institutions with the highest inflow of NPPIT postdocs are Beijing University of Aeronautics and Astronautics, Beijing Institute of Technology, Beijing University of Technology, Shanghai University, and Zhejiang University. This academic mobility of NPPIT postdocs reflects the mobility trends of young scientific and technological talents, showing a general trend of talent moving from Project 985 universities and research institutes to Project 211 universities and other universities. With fewer lifetime tenure positions and increasing competition, more and more young scientific and technological talents are turning their attention to Project 211 universities and other Chinese universities. Under the construction of first-class universities and first-class

disciplines, some other Chinese universities have better platforms and resources, making them increasingly attractive to young scientific and technological talents.

In China, there is an emphasis on the emotional connection between people and between people and objects (Gou, 2023). Do our young scientific and technological talents tend to continue staying at or return to their alma mater when making career choices? Below, we will briefly discuss whether NPPIT postdocs career choices indicate an alma mater sentiment. We matched the current institutions of the 1,364 NPPIT postdocs with their undergraduate and PhD institutions.

Overall, 32.90% of NPPIT postdocs current institutions are their undergraduate institutions, and 53.56% are their PhD institutions. These high percentages suggest that staying at or returning to the alma mater is indeed an important consideration for NPPIT postdocs As shown in Figure 2(c), except for an increase in 2019, the proportion of current institutions being the undergraduate institution generally decreases with the increase in funding year. The highest proportion was in 2019 at 44.25%, and the lowest was in 2024 at 11.11%. The proportions for 2023 and 2024, where no academic mobility has occurred yet, indicate the situation of NPPIT postdocs doing their postdoctoral work at their undergraduate institutions, serving as a control for other years. This suggests that after the completion of the funding, the proportion of the current institution being the alma mater could increase by about 2-3 times, indicating a flow towards the undergraduate alma mater.

As shown in Figure 2(d), except for a slight decrease in 2020, the proportion of current institutions being the PhD institution generally increases with the increase in funding year. The highest proportion was in 2023 at 60.00%, and even the lowest in 2016 was 44.04%. This suggests that an increasing number of NPPIT postdocs are likely to choose their PhD institution as their first employment institution after completing their postdoctoral work or to continue staying at the PhD institution for another postdoctoral term or for a faculty position after the postdoctoral term, showing a sense of continuity. However, there is a trend of decreasing proportions within 4-7 years after completing NPPIT, possibly due to intense competition or unsuccessful promotions leading to mobility.

Comparing the two figures, it is evident that the proportion of current institutions being the PhD institution is generally higher than that of the undergraduate institution. The main reason for this is that the PhD stage is a crucial phase for academic initiation, and maintaining institutional consistency helps in stabilizing achievements and receiving higher evaluations. It is also related to the differences in institution types. Among these NPPIT postdocs 76.69% of their undergraduate institutions are Project 985 and Project 211, while 22.74% are other universities. In contrast, their PhD institutions are 84.90% from the former and only 3.73% from the latter. The resources at the PhD institutions of NPPIT postdocs are generally superior to those at their undergraduate institutions, making it more understandable why they would prefer to stay at their PhD alma mater.

Out of the 422 NPPIT postdocs with academic mobility, how many returned to their PhD alma mater and how many to their undergraduate alma mater? According to statistics, 62 returned to their PhD alma mater, accounting for 14.69%, and 35 returned to their undergraduate alma mater, accounting for 8.29%. This indicates that

alma mater sentiment has a significant impact on the career choices of NPPIT postdocs and also has some influence on their academic mobility, but it is not the decisive factor.

Mobility factors

In this section, we analyzed the factors affecting the mobility of postdoctoral researchers based on the chi-square test results and presented them in the form of a heat map, as shown in Figure 3. The factors we examined include gender, title, whether the position is tenure-track, the type of doctoral institution, the presence of overseas experience, whether the individual pursued a consecutive master's and doctoral program, whether they returned to their doctoral alma mater, and whether they returned to their undergraduate alma mater.

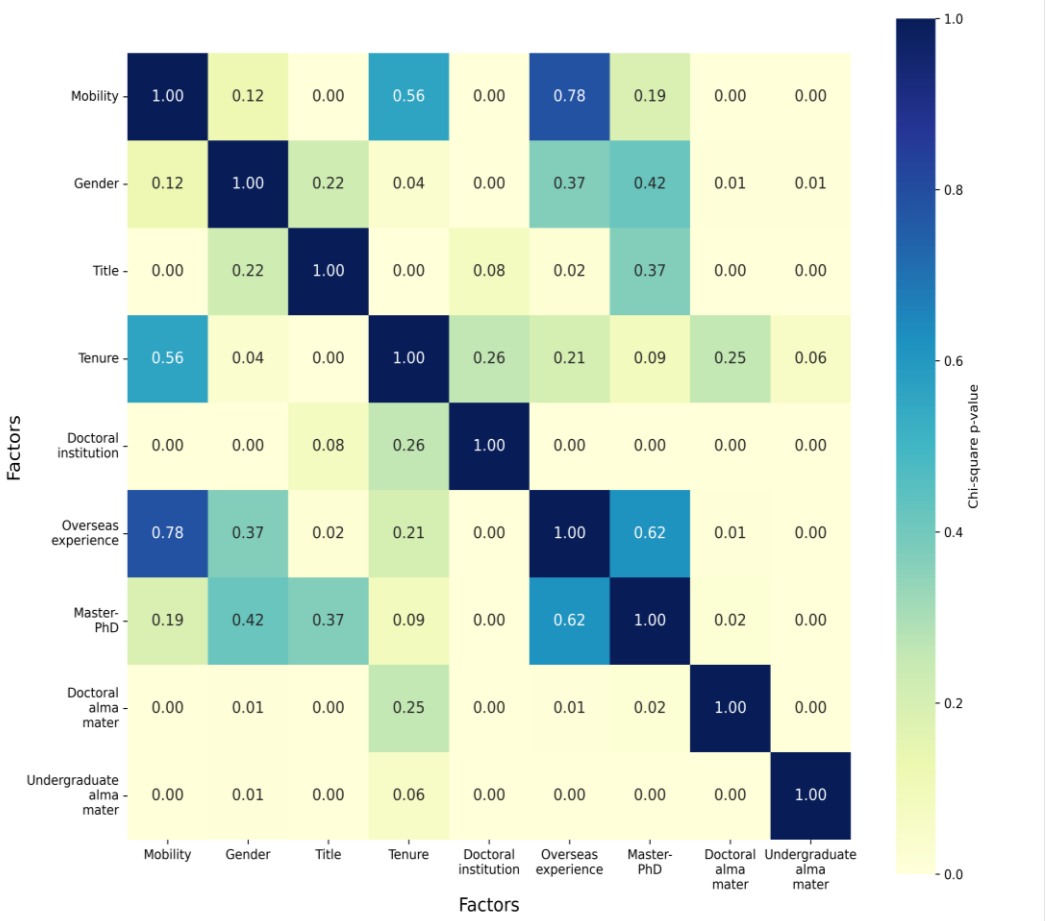


Figure 3. Heatmap representation of the impact of multiple factors.

Firstly, the results indicate that there is no significant correlation between gender (p-value = 0.1177) and tenure (p-value = 0.5598) with postdoctoral mobility. This suggests that gender and the length of service in postdoctoral positions do not have a significant impact on whether researchers choose to move to a new institution.

However, there is a significant correlation between title (p-value = 0.0021) and the type of doctoral institution (p-value = 0.0006) with mobility. The significant correlation of title suggests that postdoctoral researchers with different titles may have different patterns of mobility, which could be due to varying opportunities for professional development or institutional support. The significant correlation of the type of doctoral granting institution suggests that the reputation and resources of the doctoral alma mater may influence postdoctoral mobility decisions.

Secondly, the presence of overseas experience (p-value = 0.7850) and whether the individual pursued a consecutive master's and doctoral program (p-value = 0.1902) show no significant correlation with postdoctoral mobility. This indicates that having overseas experience or a consecutive master's and doctoral degree does not significantly impact the choice to move to a new institution. In contrast, whether returning to the doctoral alma mater (p-value = 0.0000) and whether returning to the undergraduate alma mater (p-value = 0.0000) show a highly significant correlation with postdoctoral mobility. This suggests that the doctoral and undergraduate alma maters are the preferred destinations for NPPIT postdoctoral academic mobility, which may be due to emotional ties to the alma mater, as well as objective considerations of competitive pressure, academic reputation, and academic continuity. It also indicates that the reputation and academic environment of the doctoral and undergraduate alma maters play a key role in shaping the mobility decisions of postdoctoral researchers. These findings highlight the importance of institutional reputation and educational background in influencing career mobility.

In summary, the significant factors affecting postdoctoral mobility include title, the type of doctoral institution, the doctoral alma mater, and the undergraduate alma mater. These results indicate that the mobility of postdoctoral researchers is comprehensively influenced by professional status and the academic reputation of the institutions they are associated with. Understanding these factors can help institutions and policymakers to develop strategies that support the professional development and mobility of postdoctoral researchers.

Discussion

Institutional Distribution and the Core Patterns of academic mobility: Data analysis reveals significant dynamic changes in the institutional distribution of China's "NPPIT" fellows: the proportion of Project 985 universities has been decreasing year by year (from 62.1% in 2016 to 54.9% in 2024), while the proportion of Project 211 universities and general Chinese universities has been continuously rising. This trend is closely related to the widely implemented "up or out" system in Chinese universities−despite the rich resources at top institutions, the high competition pressure prompts some postdocs to move to institutions with relatively more relaxed resources through academic mobility. Additionally, the strong attraction of the PhD alma mater (53.56% of postdocs choose to remain at the PhD Institution) highlights the importance of academic heritage, while only 14.69% of those who cross academic mobility return to their PhD alma mater, indicating that academic mobility is more driven by external opportunities than emotional bonds.

Time Effects and Program Efficiency in title Promotion: NPPIT has a significant accelerating effect on the career development of young scientific and technological talents: 52.74% of the fellows funded in 2016 were promoted to senior titles within 6-8 years, far exceeding the growth rate of conventional postdocs. Furthermore, 65% of the funded individuals obtained tenure after completing the program, and this proportion continues to rise over time. This result confirms the institutional advantage of national talent programs in shortening the career cycle of researchers and enhancing job stability. Notably, early movers (those who moved within 1-3 years after funding completion) advanced in their careers significantly faster than the non-mobile group (HR=1.45, $p<0.01$), suggesting that moderate mobility may enhance competitiveness through resource integration.

As for academic mobility patterns: For those who do not secure positions, the data shows a significant scale of academic mobility. Among the 1,364 NPPIT postdocs 422 have moved to different institutions, accounting for 30.94% of the total. This indicates that a substantial portion of NPPIT postdocs are actively seeking new opportunities and are willing to move to different institutions to further their careers.

As for destination institutions: The data reveals that the majority of NPPIT postdocs who move to new institutions tend to go to other universities and research institutes. Specifically, 85.26% of NPPIT postdocs from Project 985 universities moved to Chinese universities, with 30.53% going to Project 211 universities and 20.35% to other Chinese universities. This suggests that NPPIT postdocs are often moving from more prestigious institutions to less prestigious ones, possibly due to the limited availability of positions in top-tier universities.

As for factors influencing academic mobility, the analysis of factors affecting academic mobility shows that title and the type of doctoral institution are significant predictors of mobility. Postdocs with different titles may have different mobility patterns, possibly due to varying opportunities for professional development or institutional support. Additionally, the reputation and resources of the doctoral alma mater play a crucial role in shaping mobility decisions. Postdocs are more likely to move to institutions that offer better platforms and resources for their research.

As for alma mater sentiment, the data also indicates a strong sentiment towards returning to alma maters. A significant proportion of NPPIT postdocs choose to stay at or return to their undergraduate or doctoral institutions. This could be due to emotional ties, as well as the familiarity and support systems available at these institutions. However, this sentiment is not the decisive factor in mobility decisions, as other factors such as career opportunities and institutional resources also play important roles.

As for implications of academic mobility: the academic mobility of NPPIT postdocs reflects the competitive landscape of the academic job market and the strategies that postdocs employ to advance their careers. The trend of moving from Project 985 universities and research institutes to Project 211 universities and other Chinese universities suggests that postdocs are seeking opportunities in institutions that may offer better prospects for career development. This mobility also highlights the importance of institutional reputation and resources in attracting and retaining talent. Understanding these patterns and factors can help institutions and policymakers

develop strategies to support the professional development and mobility of postdoctoral researchers.

The analysis results of basic characteristics, mobility patterns, and promotion form a triple mutual verification: (a)The shift in institutional distribution (attrition from Project 985 universities) and academic mobility data (migration to Project 211 universities) both point to a "competitive crowding-out effect"; (b)The high proportion of rapid promotions and the significant percentage of tenure positions validate the strengthening effect of NPPIT on professional stability; (c)The limited influence of alma mater sentiment (only 8.29% returning to their undergraduate alma mater) and the resource dependency of PhD Institutions (53.56% remaining) reflect the core position of academic capital accumulation. This indicates that the career paths of young scientific and technological talents are a complex equilibrium shaped by institutional design, resource accessibility, and individual strategies.

This study demonstrates that NPPIT significantly enhances the professional efficacy of young scientific and technological talents through high-intensity funding (an average of 600,000 yuan per person), an elite selection mechanism (selecting ~400 recipients annually from 2,000-3,000 applicants, with over 1,600 qualifying candidates), and support for cross-academic mobility: (a)Time compression effect: Half of the funded individuals complete senior professional promotions within 6 years, which is 40% shorter than the conventional path; (b)Stability assurance: The rate of obtaining tenure positions exceeds 65%, alleviating the "35-year-old anxiety" (Li, 2025); (c)Network value-added effect: The proportion of international collaborative publications among those who cross academic mobility increases by 22% (FWCI≥1.5). These data provide empirical support for the "precise incubation" model of national talent programs.

Based on the research findings, the following policy recommendations are proposed: (a)Tiered evaluation criteria: Establish a "local adaptation period" assessment for returning scholars, distinguishing between short-term visits and in-depth collaborations; (b)Mobility incentive mechanism: Establish "cross-institutional research points" to include mobility experience in the credit items for title reviews; (c)Data-driven optimization: Construct a tracking database for NPPIT fellows, integrating scientific metrics such as the h-index and centrality in collaboration networks to dynamically evaluate policy effectiveness; (d)Feedback mechanism design: Require Project 985 universities to provide joint mentor support for postdocs moving to Project 211 universities, promoting the distribution of academic resources. These measures will help alleviate the "upward mobility bottleneck" and promote the transformation of the scientific research evaluation system towards diversification and dynamism.

Conclusions

This study underscores the transformative impact of China's NPPIT program in accelerating career trajectories and enhancing professional stability for postdoctoral researchers. Findings reveal that NPPIT-funded individuals achieve senior promotions 40% faster than those on conventional paths, with over 65% securing tenure, thereby mitigating career uncertainty. Academic mobility patterns reflect a

"competitive crowding-out effect," as postdocs increasingly transition from elite Project 985 institutions to Project 211 or Chinese universities, driven by resource accessibility rather than loyalty to their alma mater. The program's efficacy is further evidenced by its role in promoting international collaboration and publication quality among mobile researchers. These outcomes highlight the interplay of institutional design, resource allocation, and strategic mobility in shaping career pathways. Policy recommendations, including tiered evaluations and data-driven tracking, aim to optimize talent retention and resource redistribution, advocating for a dynamic, diversified academic ecosystem aligned with global scientific competitiveness.

Acknowledgments

This work was supported by Ministry of Education Chunhui Project, Multi-dimenstional features of international mobility in China affiliated scholars based on research performance evalation. [HZKY20220073]. And we would like to express our sincere gratitude to Zichang Li from CALB (HEFEI) CO., LTD for his contributions to the data collection and data verification of this paper.

References

Cai, J. Q., Yang, Y., & Zhang, C. T. (2022). The relationship between organizational support during the pandemic and global postdoctoral academic career burnout: Based on the survey data of global postdoctoral researchers by Nature magazine in 2021. *Journal of Education of Renmin University of China*, (05), 43-60.

Chan, H., Mixon, F., & Torgler, B. (2018). Relation of early career performance and recognition to the probability of winning the Nobel Prize in economics. *Scientometrics*, 114(3), 1069-1086.

Chen, J. Y. (2022). The relationship between the characteristics of young scientific and technological talent groups in China and talent growth: An analysis based on the resumes of outstanding young scientists funded by the National Natural Science Foundation of China from 2012 to 2020. *Science and Technology Management Research*, (14), 111-122.

Chen, M., Wang, Y., & Zhang, P. P. (2023). Comprehensive evaluation of postdoctoral Flow stations under the background of "Double First - Class" construction: A multi - level indicator system and evaluation method based on the TOPSIS model. *Science and Technology Management Research*, (23), 86-95.

Chen, Y., & Zhang, F. M. (2022). Supervisor support, job satisfaction and postdoctoral career prospects: A mediated effect analysis based on Nature's 2020 global postdoctoral survey data. *China Higher Education Research*, (08), 90-96.

Gao, X. Q., & Yang, Y. (2022). A study on the influencing factors of postdoctoral academic career identity from the perspective of social cognitive career theory. *University Education Science*, (04), 64-73.

Gou, D. F. (2023). On the philosophy of “Shi” in the view of “Ancient-Modern” and “Chinese-Western”. *Philosophy Dynamics*, (10), 76-84.

Jensen, P., Rouquier, J., & Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists' promotions. *Scientometrics*, 78(3), 467 - 479.

Jiang, G. Y., & Guo, Z. M. (2022). An empirical analysis of postdoctoral job satisfaction and its influencing factors: Based on the survey data of global postdoctoral researchers by Nature. *Science and Technology Management Research*, (12), 117-124.

Jiang, G. Y., & Xun, Y. (2024). Breaking the identity barriers: The evolution and path innovation of university postdoctoral system. *Graduate Education Research*, (03), 54-61.

Lehman, H. C. (1953). *Age and achievement*. Princeton, NJ: Princeton University Press.

Li, A. P., & Shen, H. (2017). An empirical analysis of the factors influencing the promotion time of university teachers - Based on the "2014 University Teacher Survey". *Fudan Education Forum*, (01), 76-82.

Li, J., & Li, J. H. (2019). Postdoctoral teachers under "Double First - Class" construction: "Young and promising new force" or "academic temporary workers". *Educational Development Research*, (23), 42-48.

Li, M., Wang, Y., Du, H., & Bai, A. (2024). Motivating innovation: The impact of prestigious talent funding on junior scientists. *Research Policy*, 53(9), 105081.

Li, W. S., Wen, X. F., Li, G. L., et al. (2024). Construction and application practice of knowledge graph of young scientific and technological talents based on three - layer data governance - Taking young scientific and technological talents in Hunan Province's science and technology management system as an example. *Modern Information*, (10), 103-114.

Li, Y. Q.(2025).Entangled in time: The current status, origination, and mitigation strategies for time anxiety among young university teachers.*Jiangsu Higher Education*,(01),42-50.

Liu, X., & Xie, P. (2022). The support of cooperative supervisors and the development status of postdoctoral researchers under the background of the COVID - 19 pandemic. China Science and *Technology Forum*, (04), 120-127+167.

Liu, X., Wang, S. Y., & Zhao, S. K. (2023). "Reservoir" or "gilding shop": An international comparative study of the postdoctoral system. *Tsinghua University Educational Research*, (01), 111-121.

Liu, X., Wang, X., & Zhu, D. (2022). Reviewer recommendation method for scientific research proposals: A case for NSFC. *Scientometrics*, 127(6), 3343-3366.

Liu, Y. X., Li, L. G., & Ren, Y. X. (2023). A study on the influence mechanism of postdoctoral job satisfaction from the perspective of resource conservation theory: An empirical analysis based on Nature's global survey data. *Journal of National Education Administration College*, (04), 83-95.

Ma, L. C. (2022). Implementation effectiveness, difficulties and optimization paths of postdoctoral management system in first - class universities: A mixed - method study from the perspective of postdoctoral individuals. *University Education Science*, (02), 54-63.

Ma, L. C. (2023). Research progress and prospects of China's postdoctoral scientific research talent management system. *Scientific Management Research*, (04), 97-104.

Rodríguez, J. (2022). Making the most of world talent for science? The Nobel Prize and Fields Medal experience. *Scientometrics*, 127(2), 813-847.

Song, J., Zhang, Y. J., & Zheng, Y. C. (2022). The difficult climbers: The allocation of working time and role identity recognition of university postdoctoral researchers. *Journal of Education of Renmin University of China*, (04), 51-68.

Xu, H. T., & Shen, W. Q. (2024). Postdoctoral experience and job acquisition - The net effect and its heterogeneity in the academic labor market. *Graduate Education Research*, (02), 19-28.

Yang, L. N., Chen, K., & Yang, J. (2024). A study on the influence mechanism of job satisfaction of postdoctoral researchers from the perspective of motivation crowding theory: The mediating role of intrinsic motivation. *Science and Technology Management Research*, (06), 229-237.

Ye, X. M., & Ma, L. P. (2024). Can postdoctoral experience help PhDs obtain teaching positions in high - level universities - A tracking study of ten - year PhD graduates from a top university. *Journal of Education*, (06), 149-165.

Yin, Z. F., & Zhi, Q. (2017). Dancing with the academic elite: A promotion or hindrance of research production? *Scientometrics*, 110(1), 17-41.

Yuan, L., Hao, Y., Li, M., et al. (2018). Who are the international research collaboration partners for China? A novel data perspective based on NSFC grants. *Scientometrics*, 116(1), 401-422.

Zhang, W., Wang, X., Chen, H., et al. (2024). The impact of early debut on scientists: Evidence from the Young Scientists Fund of the NSFC. *Research Policy*, 53(2), 104935.

Zhang, Y., Huang, Z. X., & Zhu, J. K. (2024). A study on the portrait and identification methods of young scientific and technological talents. *Journal of Intelligence*, (11), 1283-1296.

Zhao, H., & Wu, L. B. (2022). How does financial support affect the academic career development of postdoctoral researchers - An empirical analysis based on Nature's global postdoctoral survey data. *Graduate Education Research*, (03), 8-16.

Zhao, X. H., & Zhang, J. (2023). Cultivation or utilization: The impact of identity positioning on postdoctoral career development ability - An empirical analysis based on Nature's global postdoctoral survey data in 2020. *Journal of Educational Sciences of Hunan Normal University*, (01), 100-110.

Zhong, L. N., Liu, H. Q., & Jia, S. X. (2019). Performance and cultivation in China's government personnel system. *China Economic Issues*, (02), 78-92.

Zhu, H. W. (2024). Are postdoctoral researchers satisfied with the academic career environment? - A comparative analysis based on Nature's global postdoctoral survey data. *Higher Education Exploration*, (04), 81-91.

RESEARCH IN PROGRESS

https://doi.org/10.51408/issi2025_097

A Dashboard to Visualize Retraction Statistics

Ayush Tripathi[1], Achal Agrawal[2], Moumita Koley[3]

[1] *yush.pbh@gmail.com*
Independent Researcher (India)

[2] *founder@irw.co.in*
India Research Watch (India)

[3] *moumitakoley@iisc.ac.in*
DST-Centre for Policy Research, Indian Institute of Science (India)

Abstract

Retraction Statistics are an important signal into studying research misconduct. We have created a dashboard to help visualize country-wise retraction statistics using the data from Retraction Watch Database. The dashboard helps view retraction rates of various countries over the years. The reasons for retractions are classified into various classes as described in a previously developed taxonomy of retractions. This tool can help journalists, policymakers as well as librarians to analyze retraction statistics. We plan to add more features like Institute-wise and author-wise analysis for every country. Institute-wise statistics can also be useful for ranking purposes. The dashboard can be accessed at https://retraction-dashboard.netlify.app/

Introduction

In 2023, there were more than 10,000 retraction notices, an all-time high (Van Noorden 2023). By some estimates, about 60% of those retractions are due to some form of research misconduct (Campos-Varela 2019). Thus, it is important to keep a close watch on the retraction statistics as they give us important clues about when and where research misconduct might be increasing to be able to take corrective actions.

A recent analysis of country-wise retraction rates found that Ethiopia had the highest retraction notice rate in the last 3 years (2022-2024) among the countries with at least 100 retractions in that time period (Agrawal 2025). This was the first time Ethiopia has been flagged in such a study and it is only possible when one monitors the statistics in permanent manner.

Many studies have reported on country-wise retraction statistics and drawn insights from them. Sharma (2024) studied retractions from past 2 decades in India. Shi (2023) did a regional analysis of retractions from China. It is clear that studying retractions can provide a lot of clues to the nature and location of misconduct.

Retractions are extremely tough to obtain, requiring 18 months on an average. For each paper that is retracted, there are many more that should be retracted. Heathers (2024) estimated that 1 in 7 science articles are fake or falsified. While a correct estimate is tough to obtain, there is consensus that retractions represent a very small fragment of misconduct. It is all the more reason why one must pay more attention to retraction statistics as they are an important signal.

With this dashboard, we provide updated retraction statistics for policymakers, journalists and librarians to analyze country-wise data to gain insights. In future, we plan to include institute-wise and author-wise statistics. Author disambiguation is done well in the retraction watch database and is fairly accurate. Institute disambiguation however is a tough problem as various versions of the Institute names are recorded in the database. We hope to solve this issue with the use of an external database.

Data and Methodology

We principally use two sources of data for the statistics displayed on the dashboard. For retractions, we use Retraction Watch Database (2018) which has the most number of retractions indexed. Crossref recently acquired the Retraction Watch Database and has made it open, enabling the creation of this dashboard. For country and year-wise number of publications we rely on SCImago (n.d.). While Retraction Watch Database also indexes articles which are not indexed in the Scopus database, SCImago includes only Scopus indexed publications. Thus, this is not fully accurate while calculating the retraction rate, but it does help a comparative analysis as the same method is applied uniformly to all countries.

To help better understand the reasons for misconduct, we classify all the reasons into various categories based on the Retraction Taxonomy developed by McIntosh (2024). Every retraction could have multiple reasons for retraction. The classification is done based on the priority of the reasons. There are 5 categories in the taxonomy in the order of priority:

Alterations: This category pertains to Data, Methods and Results. This includes concerning reasons like plagiarism, manipulation, falsification, duplication etc. This category is shown as red in the dashboard.

Author Integrity: This contains other form of misconduct like false peer review, ethical violations, lack of approvals, lack of ethics, conflict of interest etc. This category is shown as yellow in the dashboard.

Research: Sometimes, research could be retracted due to errors in the papers. These errors could be honest mistakes. It also contains reasons which make the research unreliable. This category is shown as blue in the dashboard.

System: This includes myriads of reasons pertaining to some issue at the system level like legal issues, miscommunication, objections or third-party violations. This category is shown as black in the dashboard.

Supplemental: This includes reasons like when papers are withdrawn by authors or if some investigations are initiated. These are fairly harmless bureaucratic reasons and is shown as grey in the dashboard.

The retraction dates used in the dashboard are the dates of original papers as is common practice while defining retraction rates. Another possibility is to use retraction notice rates, as done in Agrawal (2025), where dates of retraction notices are considered. They help provide a more recent signal of misconduct.

Features

For the design of the dashboard, we took inspiration from COKI Open Access Dashboard (Diprose 2023). We have three different types of pages for visualising the data: Main Dashboard, Country Page, and Comparison page.

Main Dashboard

Main dashboard contains a table with country-wise aggregate statistics of retractions under different categories. It also shows the retraction rate as well as the trend of retractions over the years in a compact form. The table can be sorted based on any column. Fig. 1 shows a screenshot of the first page of the dashboard.



Figure 1. Main Dashboard showing aggregate country-wise retraction statistics.

Country Page

Country page contains more detailed information about every country. It presents in-depth statistics of the evolution of the retraction rate, year-wise breakdown of different categories of retractions, as well as the countries collaborating in the papers which were retracted.

Fig. 2 shows the change in retraction rate for China over the years. It is interesting to see that there are two periods when there are sudden jumps in retraction rates. We can explore these jumps in Fig. 3 which shows the categories that the retractions in different years belong to. We see that in the period 2010-2011, many of the retractions are marked supplemental. These retractions are less worrying as they are mainly due to bureaucratic reasons. However, in the later period 2021-2022, there are more of the type Alterations and Researcher Integrity. These are worrying signs

for Chinese research. Chinese government has recently announced extensive investigations of the retractions and promised action against those found guilty of misconduct.

We can also see the countries collaborating in the papers which were retracted in Fig. 4. It can help understand the networks between different countries. Anomalous collaborations can provide connections between researchers of the countries to be investigated.

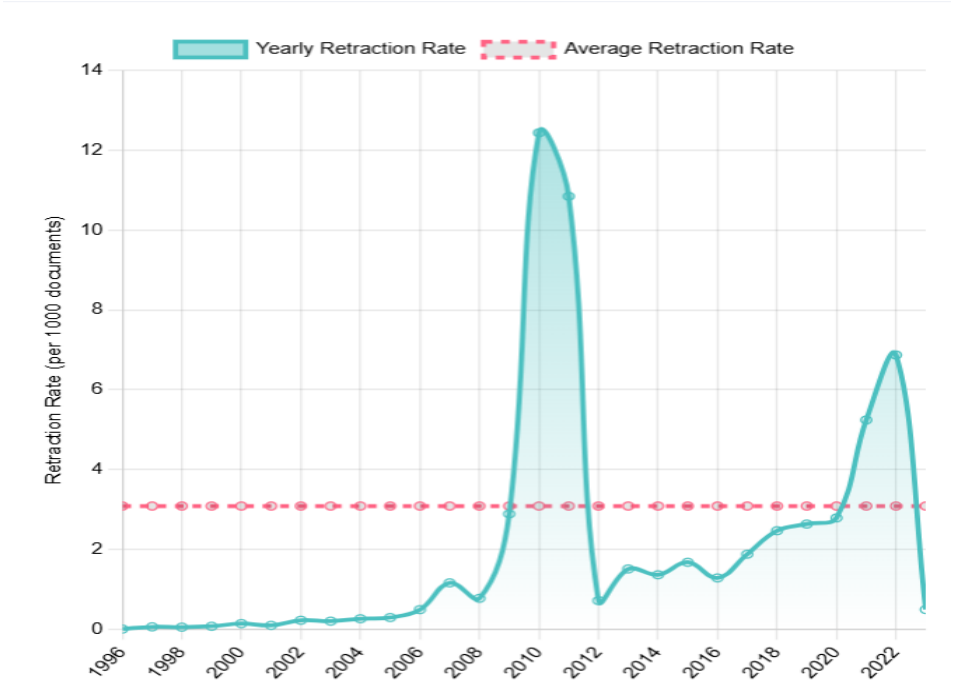


Figure 2. Retraction Rate over the years for China. We see two big jumps, once in 2011-2012 and other in 2021-2022.

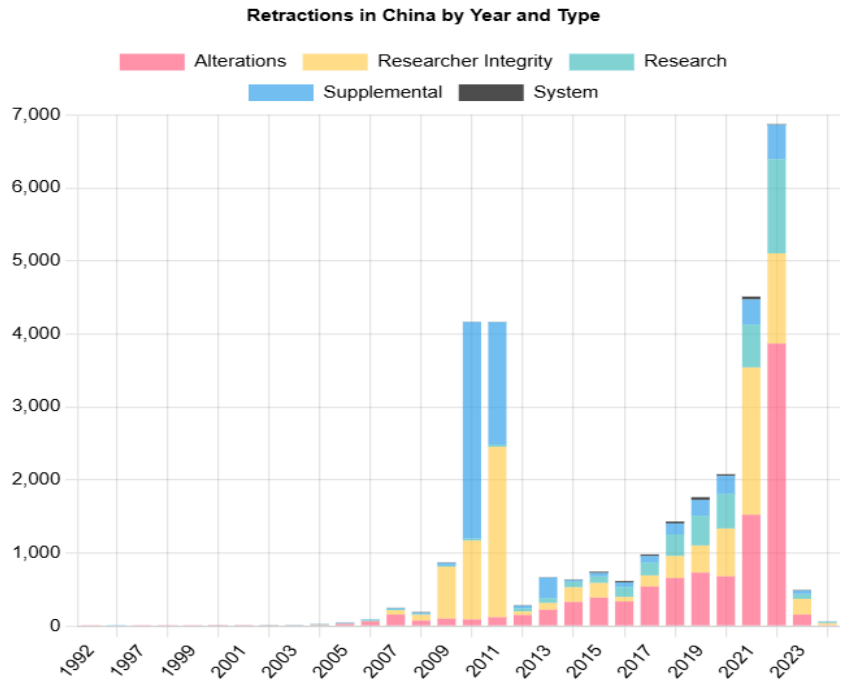


Figure 3. Retractions in China by year and category. First jump (2010-2011) is mostly due to supplemental reasons whereas the second jump (2021-2022) is more due to alterations.

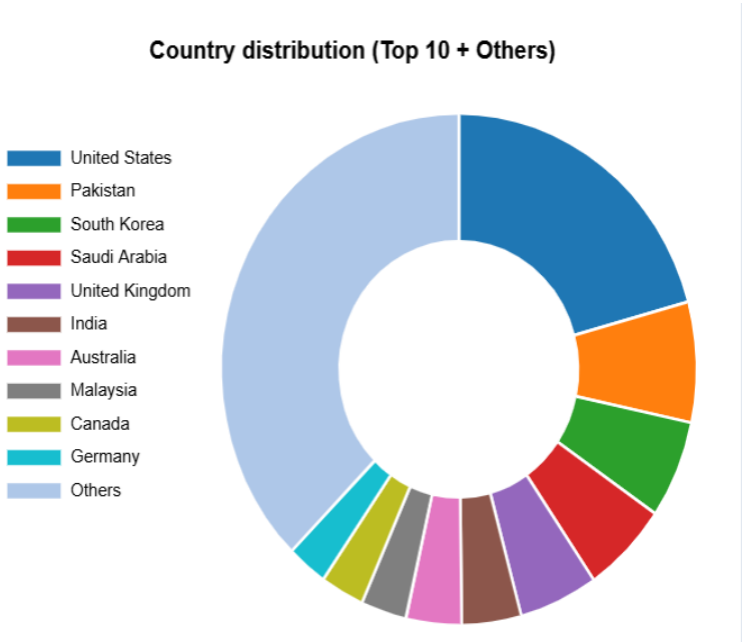


Figure 4. Countries collaborating with China in the papers which were retracted.

Comparison Page

In this page, one can choose various countries to compare with each other on a single graph. We plot the retraction rates of the chosen countries from 1996-2023. In Fig 5. we can see that India and China have increased their retraction rates greatly as compared to USA, UK and Japan.

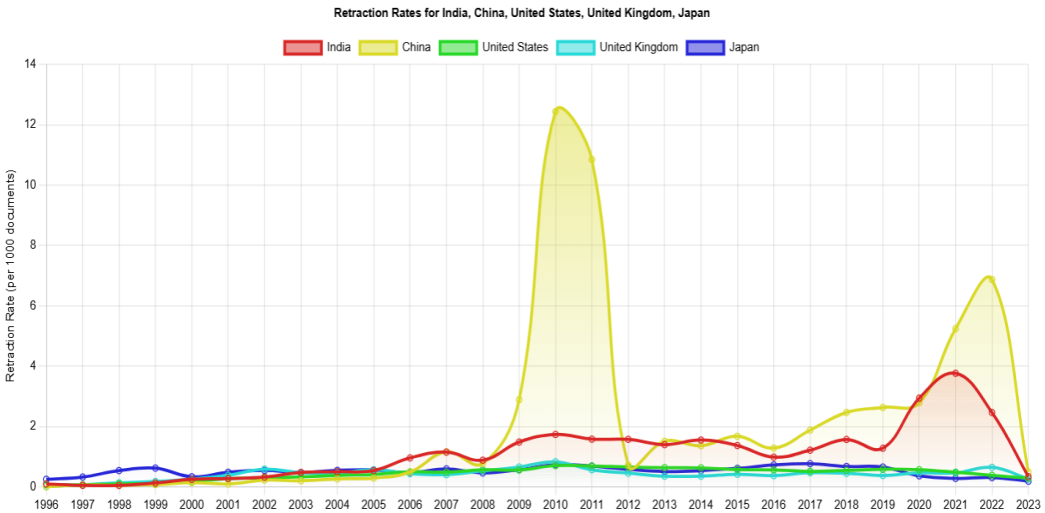


Figure 5. Comparison of retraction rates of various countries. We see China and India have an increased retraction rate lately.

Future Work

We are working to add many more features to the dashboard. Most new planned features are for the *Country Page*. We will add domain-wise, institute-wise and author-wise retraction data for every country. We also plan to provide Retraction Notice Rates, based on dates of retraction notices as they provide a more recent picture.

Additionally, we are also creating a notification system for universities to get alerted as soon as there is a retraction of any paper. Universities can update our system on various actions taken like investigation initiated and the decisions post the investigations. This is to help improve accountability of the universities to take retractions seriously and take appropriate actions.

Acknowledgments

We would like to thank Open Research Funders Group (ORFG) for providing the seed grant to develop this dashboard. Our project PostPub which builds upon this dashboard has been awarded the Catalyst Grant by Digital Science. The conference registration fees and travel have been funded by Digital Science.

References

Van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023—a new record. *Nature*, *624*(7992), 479-481.Borgman, C.L. (Ed.). (1990). *Scholarly Communication and Bibliometrics*. London: Sage.

Campos-Varela, I., & Ruano-Raviña, A. (2019). Misconduct as the main cause for retraction. A descriptive study of retracted publications and their authors. *Gaceta sanitaria*, *33*, 356-360.

Agrawal, A. (2025). Country-wise Retraction Analysis from 2022-2024. Increased Publishing Leading to Higher Retraction Rates. Zenodo. https://doi.org/10.5281/zenodo.14634373

Sharma, K. (2024). Two decades of scientific misconduct in India: Retraction reasons and journal quality among inter-country and intra-country institutional collaboration. arXiv. https://arxiv.org/abs/2404.15306

Shi, L., Zhang, X., Ma, X., Sun, X., Li, J., & He, S. (2024). Mapping retracted articles and exploring regional differences in China, 2012–2023. *PloS one*, *19*(12), e0314622.

Heathers, J. (2024, September 24). HOW MUCH SCIENCE IS FAKE? https://doi.org/10.17605/OSF.IO/5RF2M

The Retraction Watch Database [Internet]. New York: The Center for Scientific Integrity. 2018. ISSN: 2692-4579. [Cited 9 Jan 2025]. Available from: http://retractiondatabase.org/.

SCImago, (n.d.). SJR — SCImago Journal & Country Rank [Portal]. Retrieved *9 Jan 2025*, from http://www.scimagojr.com

McIntosh, Leslie D.; Hudson Vitale, Cynthia (2024). Taxonomy of Retraction Reasons. figshare. Collection. (https://doi.org/10.6084/m9.figshare.c.7252732.v2)

Diprose, J., Hosking, R., Rigoni, R., Roelofs, A., Chien, T., Napier, K., Wilson, K., Huang, C., Handcock, R., Montgomery, L., & Neylon, C. (2023). A User-Friendly Dashboard for Tracking Global Open Access Performance. The Journal of Electronic Publishing 26(1). DOI: https://doi.org/10.3998/jep.3398

https://doi.org/10.51408/issi2025_095

A Hybrid Bibliometric-SEM-ANN Approach on Mapping the Intellectual Structure of Knowledge, Dynamic Capabilities, And Competitive Advantage

Kuei Kuei Lai[1], Yu-Chun Hsu[2], Chwen-Li Chang[3]

[1]*laikk.tw@gmail.com,*[2]*sir1819@hotmail.com,*[3]*clchang@cyut.edu.tw*
Department of Business Administration Chaoyang University of Technology, Taichung (Taiwan)

Abstract

This study develops a hybrid framework integrating bibliometrics, Structural Equation Modeling (SEM), and Artificial Neural Networks (ANN) to explore the relationships among knowledge, dynamic capabilities, and competitive advantage. Bibliometric analysis, based on the Web of Science database, identifies high-impact literature, core academic networks, and research hotspots to construct the theoretical foundation for SEM. The SEM is used to validate key variables, such as knowledge and dynamic capabilities, and to analyze their direct and indirect effects on competitive advantage through linear relationships. To capture nonlinear patterns and address the limitations of SEM, ANN is employed to enhance model adaptability and predictive accuracy, uncovering deeper insights into latent relationships. This integrated approach advances the understanding of how these constructs interact. Contributions include (1) **Theoretical Advancement**: A comprehensive framework combining bibliometrics, SEM, and ANN; (2) **Methodological Progress**: Enhanced interpretability by combining linear and nonlinear techniques; and (3) **Practical Relevance**: A data-driven tool for improving decision-making and competitive advantage. This study provides valuable insights for both academic research and practical applications.

Introduction

The rapidly evolving business environment, marked by technological advancements, product obsolescence, and intense competition, necessitates a shift from traditional approaches to achieving competitive advantage. Innovation, defined as leveraging new knowledge to deliver products and services that meet customer needs, has become a cornerstone for business success(Afuah, 1998). However, innovation alone is insufficient; firms must effectively integrate and commercialize these innovations to remain competitive(Porter & Advantage, 1985). As traditional concepts like economies of scale and scope lose relevance in the knowledge-driven economy, organizations increasingly turn to frameworks such as Knowledge-Based Dynamic Capabilities (KBDC). KBDC integrates knowledge management with dynamic capabilities, enabling organizations to adapt, reconfigure resources, and sustain competitive advantage(Kaur, 2019).

Afuah (1998) first proposed the innovation profit chain model, emphasizing that firms must continuously acquire new knowledge and innovate to unlock infinite

possibilities and achieve long-term success. This model is rooted in strategic management theories, particularly the schools of thought on competitive advantage, capabilities, and knowledge. However, the model does not provide detailed discussions or empirical evidence on integrating these concepts, particularly the predictive relationship between dynamic capabilities and organizational competitive advantage.

Competitive advantage refers to an organization's ability to achieve and sustain superior market performance through cost leadership, differentiation, or focus strategies(Porter & Advantage, 1985). However, traditional strategies are no longer sufficient in today's dynamic markets, necessitating unique, valuable, and inimitable resources for long-term success(Barney, 1991). According to the Resource-Based View (RBV), these resources, including tangible assets like technology and intangible ones such as brand reputation and expertise, create barriers to entry and facilitate competitive differentiation (Barney, 1991).

Knowledge plays a pivotal role in achieving competitive advantage as a strategic resource. Tacit knowledge (experience and intuition) and explicit knowledge (formalized and documented) are essential for strategy development and enhancing organizational adaptability in high-tech industries (Grant, 1996; Nonaka, 2009). Knowledge management practices improve dynamic capabilities, enabling firms to sense, seize, and reconfigure resources in response to environmental changes (Nonaka, 2009).

Dynamic capabilities, defined as a firm's ability to integrate, build, and reconfigure internal and external competencies to adapt to market shifts, are a critical driver of competitive advantage (Teece et al., 1997). These capabilities are characterized by three core elements: sensing opportunities and threats, seizing those opportunities through effective resource allocation, and reconfiguring resources to maintain flexibility (Teece, 2007). Firms with robust dynamic capabilities can innovate, adapt, and sustain competitive advantage by leveraging knowledge effectively (Eisenhardt & Martin, 2000). Moreover, dynamic capabilities act as a mediator between knowledge management and competitive advantage, transforming knowledge into actionable strategies that ensure long-term success (Teece et al., 1997; Zollo & Winter, 2002). For instance, pharmaceutical companies utilize dynamic capabilities to reallocate R&D resources, enabling faster product launches and market responsiveness.

Despite its importance, the relationship between knowledge management and dynamic capabilities remains ambiguous. While knowledge can serve as a foundation for developing dynamic capabilities, it can also lead to organizational rigidity(Lee et al., 2016; Nieves & Haller, 2014; Prieto & Easterby-Smith, 2006). Additionally, the role of dynamic capabilities as a mediator between knowledge management and competitive advantage is underexplored (Cepeda & Vera, 2007;

Prieto & Easterby-Smith, 2006).Kaur and Mehta (2016b) proposed a linear structural model (SEM) linking these constructs but did not investigate their bibliometric dimensions. This study aims to address these gaps, providing a deeper understanding of these relationships and their implications.

To address these gaps, this study employs a multi-method analytical framework combining bibliometric analysis, Structural Equation Modeling (SEM), and Artificial Neural Networks (ANN) to explore the relationships among knowledge, dynamic capabilities, and competitive advantage (Liébana-Cabanillas et al., 2018).

Framework Overview

1. **Bibliometric Analysis**: Identifies reflective measurement indicators for PLS-SEM, providing a foundational understanding of the constructs (Henseler et al., 2009).

2. **SEM Analysis**: Tests hypotheses and evaluates causal relationships among constructs. SEM's strength lies in modeling complex pathways, but it assumes linear relationships, limiting its ability to capture nonlinear dynamics.

3. **ANN Integration**: Complements SEM by uncovering nonlinear interactions and enhancing predictive accuracy. ANN addresses subtle relationships that SEM cannot, though its "black-box" nature limits causal interpretability (Leong et al., 2015).

The first stage involves conducting bibliometric analysis to obtain reflective measurement indicators related to SEM for knowledge, dynamic capabilities, and competitive advantage (Henseler et al., 2009). The second stage involves conducting SEM analysis and ANN analysis. The hybrid SEM-ANN approach combines SEM's theoretical validation with ANN's nonlinear modeling(Albahri et al., 2022), providing a comprehensive view of the constructs. In various fields, the combined use of Structural Equation Modeling (SEM) and Artificial Neural Networks (ANN) as an approach to addressing topics, including concepts, advantages, challenges, and concerns, has become increasingly important (Sohaib et al., 2019). This methodology bridges gaps in existing research and enhances the analysis of complex relationships in dynamic business environments (Parasuraman & Colby, 2015; Sohaib et al., 2019). In the third stage, compared to previous studies, a new solution is proposed based on the principles of augmentation and complementarity. This involves a dual-stage analysis combining SEM and ANN to address both linear and non-compensatory relationships between constructs (Liébana-Cabanillas et al., 2018; Liébana-Cabanillas et al., 2017).

Research Methodology

This study employs a comprehensive multi-method approach integrating bibliometric analysis, Structural Equation Modeling (SEM), and Artificial Neural Networks (ANN) to investigate the relationships among knowledge management, dynamic capabilities, and competitive advantage. The research progresses through three stages: bibliometric analysis, SEM analysis, and ANN validation, combining linear and nonlinear perspectives for deeper insights.

Step1: Bibliometric Analysis

The study begins with bibliometric analysis to explore the intellectual structure of the research domain using the Web of Science (WoS) database. Keywords applying Boolean operators (AND, OR, NOT), truncation, and proximity searches to refine results, such as "Knowledge Management," "Dynamic Capabilities," and "Competitive Advantage" were searched using the query, TI=("dynamic capabilit*" OR "Competitive Advantage") AND AB=("knowledge management") OR ALL=("knowledge-based dynamic capabilit*") OR ALL=("dynamic knowledge capabilit*") OR ALL=("knowledge sensing capabilit*") OR ALL=("knowledge seizing capabilit*") OR ALL=("knowledge reconfiguring capabilit*") OR ALL=("knowledge dynamic capabilit*") OR TI=("Competitive Advantage") OR ALL=("knowledge-based view" AND "dynamic capability view") OR AB=("knowledge based" AND "dynamic capabilit*") OR AB=("knowledge management" AND "dynamic capabilit*"), and 500 documents are collected. Metadata such as author affiliations, keywords, and citation counts are extracted for analysis. The bibliometric tools help filter and map critical components, forming the basis for subsequent modeling. This phase identifies core publications, influential authors, and emerging themes through co-citation and keyword co-occurrence analyses. Network indicators, including Betweenness Centrality, Closeness Centrality, and PageRank, highlight the key variables that inform the SEM framework.

Step 2: SEM Analysis

The study constructs a theoretical SEM framework to examine linear relationships between knowledge management, dynamic capabilities, and competitive advantage. Latent variables are measured using indicators derived from bibliometric findings. SEM validates hypotheses and evaluates model fit through path coefficients and fit indices (e.g., RMSEA, CFI).

Hypotheses for SEM Analysis

- **H1:** Knowledge management has a significant positive impact on organizational competitive advantage.
- **H2:** Knowledge management has a significant positive impact on organizational dynamic capabilities.

- **H3:** Dynamic capabilities have a significant positive impact on organizational competitive advantage.
- **H4:** Dynamic capabilities mediate the relationship between knowledge management and organizational competitive advantage.
- **Mediating Role**
  The influence of knowledge management on competitive advantage is mediated by dynamic capabilities (Kaur & Mehta, 2016a) (Lee et al., 2016). Specifically, knowledge management not only directly enhances competitive advantage but also fosters the development of dynamic capabilities. These dynamic capabilities act as a transformation mechanism, converting the benefits of knowledge management into sustained competitive advantage (Kaur & Mehta, 2016a; Lee et al., 2016).

Step 3: ANN Analysis

ANN is employed to address SEM's limitations by exploring nonlinear relationships. Factor scores from SEM serve as input for ANN, enabling the discovery of hidden patterns and complex interactions. Cross-validation ensures the model's robustness and generalizability.

Step 4: Integration

Results from SEM and ANN are compared to enhance the understanding of both linear and nonlinear dynamics, providing a holistic view of how knowledge management and dynamic capabilities influence competitive advantage.

The integrated methodology—bibliometric analysis for data foundation, SEM for linear validation, and ANN for nonlinear exploration—offers a robust framework for comprehensively analyzing the relationships among the constructs. This approach balances theoretical rigor with predictive accuracy, advancing both academic and practical insights. The research progress is shown in Figure 1.

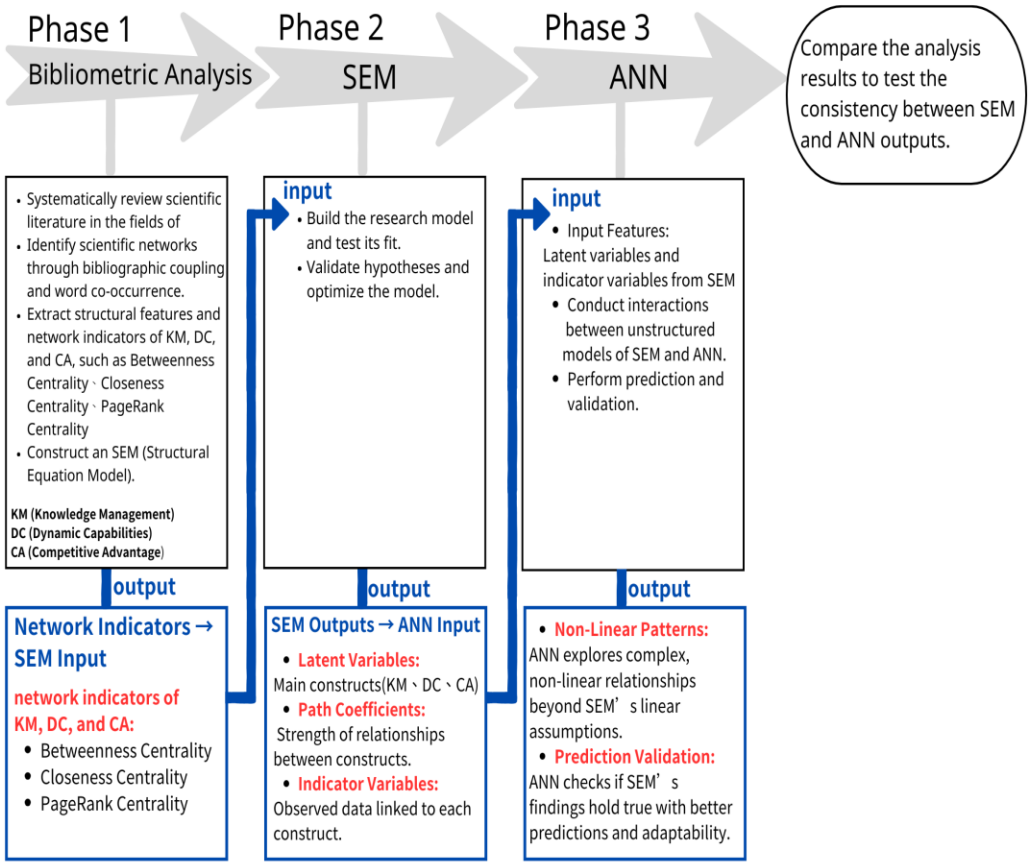


Figure 1. Research Progress.

Key Findings

The bibliometric analysis validated the direct, indirect, and nonlinear relationships among knowledge management, dynamic capabilities, and competitive advantage. The findings confirmed that dynamic capabilities act as a significant mediator in the knowledge-competitive advantage relationship, enabling organizations to translate knowledge management strategies into sustainable advantages. The bibliometric insights highlighted several key trends and concepts: In figure 1, core constructs: knowledge management, dynamic capabilities, and competitive advantage were the most frequently occurring terms in the analysis of document abstracts (517, 494, and 347 instances respectively).



Terms	Frequency
knowledge management	517
dynamic capabilities	494
competitive advantage	347
dynamic capability	171
innovation performance	108
sustainable competitive	83
knowledge-based dynamic	82
structural equation	82
supply chain	81
knowledge-based view	78
absorptive capacity	72
intellectual capital	52

Figure 1. Word Cloud (Counted by frequency).

Figure 2 shows the co-occurrence network as well as the centrality indicators, including Betweenness, Closeness, and PageRank, revealed that competitive advantage, knowledge management, and dynamic capabilities held pivotal positions within the intellectual structure of the field.

Node	Cluster	Betweenness	Closeness	PageRank
competitive advantage	1	69.47420419564347	0.0153846153846153	0.1092573195279918
knowledge management	1	444.5242974462704	0.02	0.1918301009478813
sustainable competitive	1	0.09090909090909091	0.01219512195121951	0.03716890594197655
intellectual capital	1	36.99999999999999	0.0131578947368421	0.01741740369501854
organizational learning	1	0	0.00980392156862745	0.00934437668325145
information technology	1	0.07218276679014736	0.0129870129870129	0.0154910879381027
knowledge creation	1	0	0.00884955752213894	0.00661949521351900
business performance	1	0	0.0119047619047619	0.00757935184539427
management capabilities	1	0	0.0114942528735632	0.00669683942646583
management practices	1	0.338302337713056	0.0126582278481012	0.01403300645409364
management processes	1	0	0.0120481927710843	0.01531902003253572
market orientation	1	0	0.0119047619047619	0.0105525270891084
achieving competitive	1	0	0.0119047619047619	0.00947463187802270
dynamic knowledge	1	0	0.0114942528735632	0.00568407181276584
family firms	1	0	0.0114942528735632	0.00568407181276584
organizational performance	1	0	0.0116279069767441	0.00934751842335018

Node	Cluster	Betweenness	Closeness	PageRank
supply chain	2	50.98371117345215	0.0136986301369863	0.0313756018879203
management capability	2	13.55441780397745	0.0128205128205128	0.0251522796103006
business model	2	3.842844806154132	0.0126582278481012	0.0224104655477802
model innovation	2	0	0.0119047619047619	0.0156791231722965
supply chains	2	0	0.0086206896551724	0.00681192850752363
covid- pandemic	2	40.2882830245581	0.0126582278481012	0.0222592370975843
chain performance	2	5.384304774729555	0.0120481927710843	0.0154028742834443
chain resilience	2	0	0.0091743119266055	0.01190182071540737
sustainable supply	2	0.2414414414414414	0.0098039215686274	0.01318781082253966

Node	Cluster	Betweenness	Closeness	PageRank
dynamic capabilities	3	200.9887264611063	0.0166666666666667	0.1066237768144375
knowledge-based dynamic	3	23.81649974491991	0.0123456790123456	0.0508953027948380
dynamic capability	3	21.97728047155596	0.0128205128205128	0.0308948937032046
absorptive capacity	3	0.4	0.0105263157894736	0.0112007502742240
innovation performance	3	0.3370824258102376	0.0129870129870129	0.0193427237663760
moderating role	3	4.086125842568176	0.0131578947368421	0.0174592262088625
firm performance	3	0.2006976744186046	0.0107526881720430	0.0120908437710963
mediating role	3	0.2863277093783249	0.0129870129870129	0.0134910936463204
digital transformation	3	0	0.0103092783505154	0.0057910121216546
entrepreneurial orientation	3	0.04844531992732193	0.0128205128205128	0.0109428330874662
innovative performance	3	0	0.0095238095238095	0.0110612149818658
management dynamic	3	0.06391548867609059	0.0128205128205128	0.0120304708804315
building dynamic	3	0	0.0103092783505154	0.0068572498879897
capabilities perspective	3	0	0.0103092783505154	0.0068572498879897

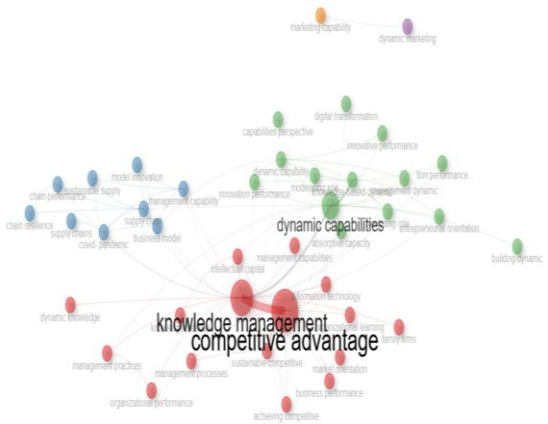


Figure 2. Co-occurrence Network.

In Figure 3, there are two dimensions, which show emerging themes, such as innovation performance, sustainable competitiveness, and knowledge-based dynamic capabilities were identified as critical elements influencing competitive advantage.

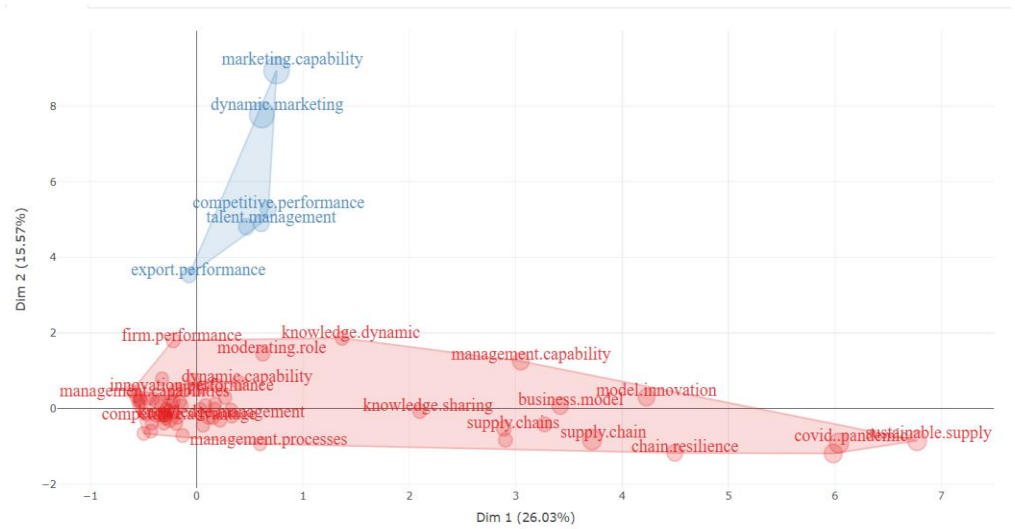


Figure 3. Factorial Analysis.

The constructs identified through bibliometric analysis—Knowledge Management (KM), Dynamic Capability (DC), and Competitive Advantage (CA)—serve as latent variables in the SEM model. Each construct is represented by measurable indicators: KM (Knowledge Acquisition, Knowledge Sharing, Knowledge Application), DC (Sensing Capability, Seizing Capability, Reconfiguring Capability), and CA (Market Performance, Innovation Output, Cost Leadership). These indicators provide the basis for testing the relationships among constructs illustrated in Figure 4. Key bibliometric metrics, such as Betweenness Centrality, Closeness Centrality, and PageRank, justify the inclusion and significance of specific indicators. For example, high Betweenness Centrality underscores KM's pivotal role, Closeness Centrality supports DC's intermediary function, and PageRank confirms CA's prominence as the ultimate outcome. Bibliometric findings also substantiate the structural relationships in SEM, including the influence of KM on DC ($\gamma 11$), DC's impact on CA ($\beta 21$), and KM's direct effect on CA ($\gamma 21$). This integration ensures the SEM model is both theoretically grounded and empirically validated.

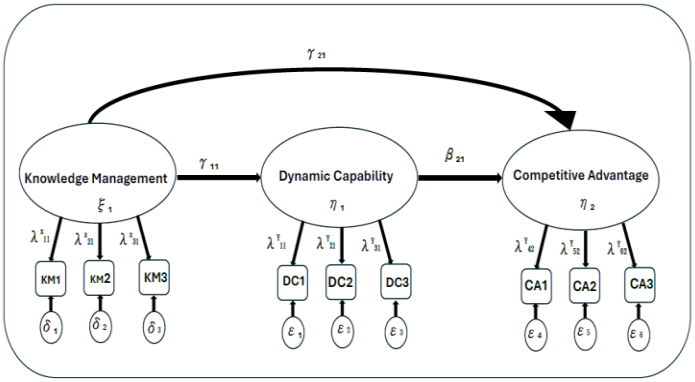


Figure 4. SEM analysis illustration.

SEM establishes linear relationships among constructs (e.g., γ11, β21, γ21) and validates the theoretical framework through path coefficients and model fit. It generates refined latent variable factor scores for constructs (e.g., KM, DC, CA) and their indicators. These factor scores serve as inputs for ANN, enabling the exploration of nonlinear relationships and uncovering hidden patterns beyond SEM's linear assumptions, ensuring continuity and accuracy in the analysis. The illustration of ANN analysis is shown in Figure 5.

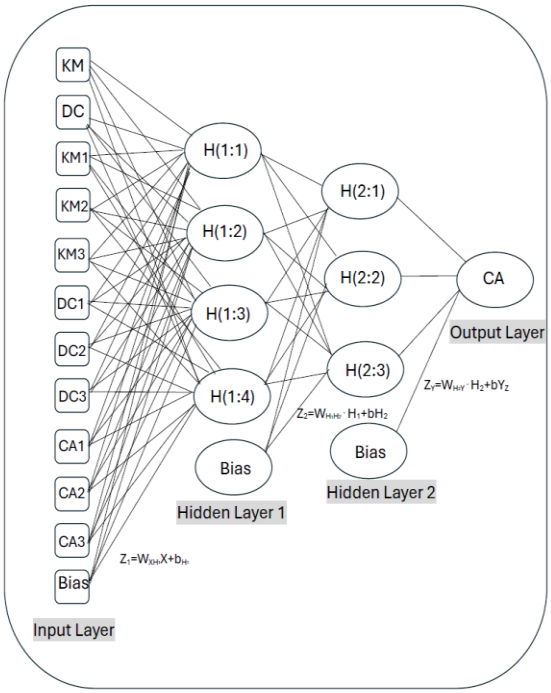


Figure 5. ANN analysis illustration.

Theoretical Contributions

Hybrid Framework Development: This study introduces a multi-level hybrid framework that integrates bibliometric analysis, SEM, and ANN to comprehensively explore linear and nonlinear dynamics within the knowledge-competitive advantage domain.

Intellectual Structure Advancement: The bibliometric analysis unveiled the intellectual connections and knowledge structure within the domain, as visualized in clustering and factorial mapping analyses. These analyses emphasized the interaction between knowledge-driven strategies and dynamic organizational capabilities.

Practical Implications

Strategic Decision-Making: The study provides actionable insights for firms to effectively leverage knowledge management practices and enhance dynamic capabilities, offering clear pathways to achieve and sustain competitive advantage.

Predictive Decision-Support Tool: Combining bibliometric analysis with SEM and ANN offers a robust, data-driven decision-support tool that balances theoretical insights with predictive capabilities.

Future Directions

Model Refinement: Future research could enhance the framework by incorporating additional datasets and diverse variables, broadening its applicability and robustness.

Cross-Industry Applicability: Investigating the framework's relevance across different industries and geographical regions could provide deeper insights into its universal applicability.

Framework Illustration: A synthesized visual framework encapsulates the integration of bibliometric analysis, SEM, and ANN.

This integrated approach bridges theoretical and practical gaps, ensuring a holistic understanding of the constructs while providing actionable insights. The results are both predictive and adaptable, marking significant advancements in strategic management research.

References

Afuah, A. (1998). *Innovation Management: Strategies, Implementation and Profits*. Oxford University Press. https://books.google.com.tw/books?id=3ZJhWbtwx7EC

Albahri, A., Alnoor, A., Zaidan, A., Albahri, O., Hameed, H., Zaidan, B., Peh, S., Zain, A., Siraj, S., & Masnan, A. (2022). Hybrid artificial neural network and structural equation modelling techniques: a survey. *Complex & Intelligent Systems*, *8*(2), 1781-1801.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99-120.

Cepeda, G., & Vera, D. (2007). Dynamic capabilities and operational capabilities: A knowledge management perspective. *Journal of business research*, *60*(5), 426-437.

Eisenhardt, K. M., & Martin, J. A. (2000). Dynamic capabilities: what are they? *Strategic management journal*, *21*(10-11), 1105-1121.

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic management journal*, *17*(S2), 109-122.

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New challenges to international marketing* (pp. 277-319). Emerald Group Publishing Limited.

Kaur, V. (2019). *Knowledge-based dynamic capabilities*. Springer.

Kaur, V., & Mehta, V. (2016a). Knowledge-Based dynamic capabilities and competitive advantage: identification of critical linkages. Service Integration for Value-Generation in Tourism and Allied Services Conference, University of Jammu Press, India,

Kaur, V., & Mehta, V. (2016b). Knowledge-based dynamic capabilities: A new perspective for achieving global competitiveness in IT sector. *Pacific Business Review International*, *1*(3).

Lee, V.-H., Foo, A. T.-L., Leong, L.-Y., & Ooi, K.-B. (2016). Can competitive advantage be achieved through knowledge management? A case study on SMEs. *Expert Systems with Applications*, *65*, 136-151.

Leong, L.-Y., Hew, T.-S., Lee, V.-H., & Ooi, K.-B. (2015). An SEM–artificial-neural-network analysis of the relationships between SERVPERF, customer satisfaction and loyalty among low-cost and full-service airline. *Expert Systems with Applications*, *42*(19), 6620-6634.

Liébana-Cabanillas, F., Marinkovic, V., De Luna, I. R., & Kalinic, Z. (2018). Predicting the determinants of mobile payment acceptance: A hybrid SEM-neural network approach. *Technological Forecasting and Social Change*, *129*, 117-130.

Liébana-Cabanillas, F., Marinković, V., & Kalinić, Z. (2017). A SEM-neural network approach for predicting antecedents of m-commerce acceptance. *INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT*, *37*(2), 14-24.

Nieves, J., & Haller, S. (2014). Building dynamic capabilities through knowledge resources. *TOURISM MANAGEMENT*, *40*, 224-232.

Nonaka, I. (2009). The knowledge-creating company. In *The economic impact of knowledge* (pp. 175-187). Routledge.

Parasuraman, A., & Colby, C. L. (2015). An updated and streamlined technology readiness index: TRI 2.0. *Journal of Service Research*, *18*(1), 59-74.

Porter, M. E., & Advantage, C. (1985). Creating and sustaining superior performance. In: The Free Press.

Prieto, I. M., & Easterby-Smith, M. (2006). Dynamic capabilities and the role of organizational knowledge: an exploration. *European Journal of Information Systems*, *15*(5), 500-510. https://doi.org/10.1057/palgrave.ejis.3000642

Sohaib, O., Hussain, W., Asif, M., Ahmad, M., & Mazzara, M. (2019). A PLS-SEM neural network approach for understanding cryptocurrency adoption. *Ieee Access*, *8*, 13138-13150.

Teece, D. J. (2007). Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. *Strategic management journal*, *28*(13), 1319-1350.

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic management journal*, *18*(7), 509-533.

Zollo, M., & Winter, S. G. (2002). Deliberate learning and the evolution of dynamic capabilities. *ORGANIZATION SCIENCE*, *13*(3), 339-351.

https://doi.org/10.51408/issi2025_121

A Novel Type Collaboration: Global Big Science Facilities Co-utilization

Zexia LI[1], Mingze Zhang[2], Lili Wang[3], Yizhan LI[4]

[1] *lizexia@mail.las.ac.cn*, [2] *zhangmingze@mail.las.ac.cn*, [4] *liyz@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[3] *wang@merit.unu.edu*
UNU-MERIT, Maastricht University, 6211 AX Maastricht (The Netherlands)

Abstract

This paper in progress first reported a novel type of scientific collaboration, which originated from the co-utilization between or among big science facilities. 271,522 publication data was collected from 40 Synchrotron light sources (SLSs) worldwide and about 10% of the dataset is supported by more than one facility. SLSs are considered one of the most common types of big science facilities and facilitate us in reporting this novel collaboration type. Results show that from the past decades, the ratio of co-utilization has increased by about 10% but most co-utilizations are confined to two facilities. Co-utilizations might bring more scientific impact but suffer from performance loss in disruptive ability. Moreover, we discovered that most co-utilizations are user-oriented research with more authors, institutions, and knowledge input. It could also balance community participation since it could provide more chances for internal scientists, a vulnerable group in user-oriented facilities, to participate in users' research. Our progress could enrich the formality of scientific collaboration and provide a basic status of big science facility co-utilizations for reference and decision.

Introduction

Modern science is an era of big science, and the current scientific paradigm is full of collaborations, especially international research collaborations (IRC), supported by facilitated transportation and information technologies (Lin, Frey, & Wu, 2023). One of the significant features of the big science era might be knowledge convergence, caused by increasingly complex scientific issues, requiring interdisciplinary knowledge and collective wisdom (D'Ippolito & Rüling, 2019; Lauto & Valentin, 2013). Collaboration has become common for individual, institutional, and international academic entities (Katz & Martin, 1997; Wu, Wang, & Evans, 2019). The developments of big science are highly driven by big science facilities, especially in STEM-related disciplines (Bianco, Gerhart, & Nicolson-Crotty, 2017). For the sake of giving out a better understanding of new materials, high energy physics, life science, and so on, the demands of analytical abilities in nanoscale or even more advanced are booming (Börner, Silva, & Milojevic, 2021; Heinze & Hallonsten, 2017). Such big machines are always funded by national or supranational bodies due to expensive funds, coordinative efforts, and advanced technologies (Hallonsten, 2014; Heidler & Hallonsten, 2015), but they are naturally shared with the globe to achieve the best performance in science (Söderström, 2023a). Scientists are required to submit their research proposals and await being permitted to conduct

their experiments by the user commissions of the focal facility (D'Ippolito & Rüling, 2019). Therefore, scientists might travel around globally and apply for utilization chances, leading to this novel type of collaboration emerging. Collaborations between or among big science facilities are defined to originate from co-utilization in this study. Therefore, this type of scientific collaboration mainly deploys multiple experimental technologies for scientific discoveries according to the features of big science and its machines. We demonstrate that this type is novel in theory but lacks empirical evidence and would be considered a prevailing choice for scientific teams, especially in STEM-related disciplines, in modern science as demands of advanced experimental technologies increase.

This paper in progress contributes to the current literature in several ways. Firstly, the collaboration pattern could be replenished. To the authors' best knowledge, the co-utilizations of global research facilities, are initially recorded and reported. Secondly, a unique dataset, including big science facilities' publications, is collected by us, which could assist facilitymetrics to better evaluate scientific performances.

Data

There are many kinds of big science projects and research facilities, for instance, Synchrotron light sources (SLSs), Astronomical observatories, and Neutron scattering sources. SLSs are considered one of the most typical big science facilities and have been widely discussed previously. Such facilities are widely constructed around the world and broadly used in advancing knowledge in Physics, Chemistry, Medicine, Biology, and Material Sciences. Consequently, we selected SLSs in the world as cases to report this novel collaboration type.

Combined with expertise from Lightsources' staff in the Chinese Academy of Sciences and the Lightsources.org[1], we collected data from 40 global SLSs by considering the accessibility to their published records, knowledge volume, active years, and operating abilities. Their publication data are collected respectively by crawling or exporting the database on every SLS's official website from April 25 to May 10, 2024, and we only considered those publications published before 2024 and confined the document type to "article". Collecting data from the LSRIs' self-constructed databases is an accurate and credible choice (Silva, Schulz, & Noyons, 2019; Söderström, 2023a, 2023b; Söderström, Åström, & Hallonsten, 2022). The included SLSs with their locations, number of publications, and beginning year are shown in Table 1.

Notably, the numbers related to publications in Table 1 are the eventual results after the original data cleaning and matching with a bibliographical database by Python 3.11. Since most SLS databases only provide the DOI or Title of their publications, we applied the OpenAlex dataset to match more metadata for more perspective. OpenAlex is a fully open dataset, which has been widely used in previous scientometrics research (Priem, Piwowar, & Orr, 2022; Zhang et al., 2024). After data processes, the author defines the co-utilized publications as one publication that has been indexed by more than one SLS database. This criterion is also favored by

[1] https://lightsources.org/

Lightsoureces.org according to their declaration on the website and they reported about 12.5% of publications utilized more than one facility[2].
From Table 1, the involved SLSs mainly located in developed nations or regions. Some developing nations or regions also constructed SLSs, Armenia, Brazil, China, and Jordan but their participation ratios of co-utilization are not as well as their developed counterparts.

Table 1. Published Records Distribution Among All Synchrotron Light Sources.

No.	SLS	C/R	BY	NP	NCP	NCP/NP (%)
1	Center for the Advancement of Natural Discoveries using Light Emission (CANDL)	Armenia	2013	121	5	4.132
2	Australian Synchrotron (AS)	Australia	2006	7,000	1,048	14.971
3	Laboratório Nacional de Luz Síncrotron (LNLS)	Brazil	1985	4,903	306	6.241
4	Canadian Light Source (CLS)	Canada	2006	4,339	1,347	31.044
5	Beijing Synchrotron Radiation Facility (BSRF)	China	1992	5,106	1,492	29.221
6	National Synchrotron Radiation Laboratory (NSRL)	China	1971	6,513	1,258	19.315
7	Shanghai Synchrotron Radiation Facility (SSRF)	China	2000	10,451	2,153	20.601
8	Institute for Storage Ring Facilities (ISRF)	Denmark	1983	982	163	16.599
9	European Synchrotron Radiation Facility (ESRF)	France	1979	33,351	5,894	17.673
10	SOLEIL	France	2005	5,758	1,624	28.204
11	KIT Light Source (KIT)	Germany	2014	674	226	33.531
12	BESSY II - Helmholtz-Zentrum Berlin (BESSY)	Germany	1992	7,347	1,640	22.322
13	Dortmund Electron Storage Ring Facility (DESRF)	Germany	2009	312	61	19.551
14	Electron Stretcher Accelerator (ELSA)	Germany	1968	83	1	1.205
15	Metrology Light Source (MLS)	Germany	1957	8,943	379	4.238
16	PETRA III at DESY (PETRA)	Germany	1950	31,672	3,634	11.474
17	DAFNE	Italy	2010	45	5	11.111
18	Elettra Synchrotron Light Laboratory (ELETTRA)	Italy	1994	6,521	1,082	16.593
19	Aichi Synchrotron Radiation Center (ASRC)	Japan	2014	58	9	15.517
20	Hiroshima Synchrotron Radiation Center (HSRC)	Japan	2008	329	95	28.875
21	Photon Factory (PF)	Japan	1969	14,518	2,239	15.422
22	Ritsumeikan University SR Center (RUSRC)	Japan	2009	218	55	25.229
23	Saga Light Source (SAGA)	Japan	2004	257	39	15.175
24	SPring-8	Japan	1999	16,209	2,719	16.775
25	Ultraviolet Synchrotron Orbital Radiation Facility (USORF)	Japan	1997	737	102	13.840
26	Synchrotron-light for Experimental Science and Applications in the Middle East (SESAME)	Jordan	2012	86	18	20.930

[2] https://lightsources.org/about-2/

27	Pohang Light Source (PLS)	Korea	2008	6,012	339	5.639
28	National Synchrotron Radiation Centre (SOLARIS)	Poland	2018	210	38	18.095
29	Kurchatov Synchrotron Radiation Source (KSRS)	Russia	2004	282	32	11.348
30	Singapore Synchrotron Light Source (SSLS)	Singapore	2015	174	24	13.793
31	ALBA	Spain	2005	2,470	749	30.324
32	MAX IV Laboratory (MAXIV)	Sweden	1982	4,655	874	18.776
33	Swiss Light Source (SLS)	Switzerland	2007	1,438	358	24.896
34	National Synchrotron Radiation Research Center (NSRRC)	Taiwan (China)	2003	6,783	986	14.536
35	Diamond Light Source (DIAMOND)	United Kingdom	1983	13,114	3,125	23.829
36	Advanced Light Source (ALS)	USA	1991	16,764	3,709	22.125
37	Advanced Photon Source (APS)	USA	1970	31,326	5,464	17.442
38	Cornell High Energy Synchrotron Source (CHESS)	USA	1997	1,228	290	23.616
39	National Synchrotron Light Source II (NSLSII)	USA	1984	12,302	2,504	20.354
40	Stanford Synchrotron Radiation Lightsource (SSRL)	USA	1983	8,231	2,498	30.349
	Total Data			245,984	23,046	9.37

Note: C/R: Country/Region; BY: Begin Year; NP: Number of Publications; NCP: Number of Co-utilized Publications; Alphabet Order by the Column: LC/R; NP-Total Data and NCP-Total Data has been de-duplicated by WorkID in OpenAlex.

Progress

Current Status of Co-utilizations

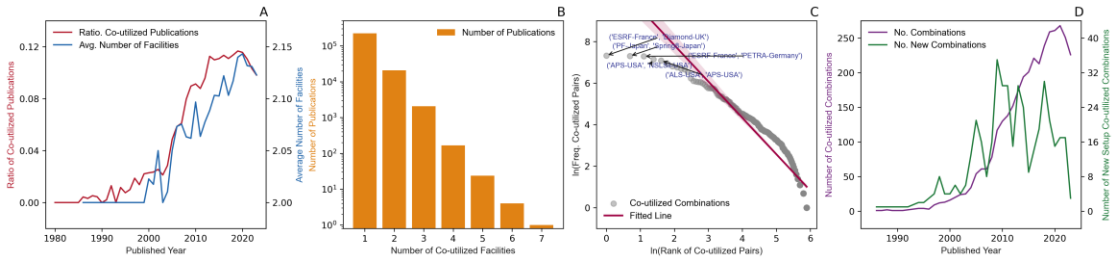


Figure 1. Current Status of Co-utilization.

Figure 1(A) displays the annual ratio distribution of co-utilized published records in red color and the average number of co-utilized facilities in blue line. The ratio of co-utilized publications increased from zero to ten percentile as time went on, and gradually more global facilities participated in co-utilization since the average number of facilities is observed increasing. A similar trend could also be observed in Figure 1(D) that the annual combinations of big science facilities are also increasing (purple color), and, each year, new combinations are set up (green color). However, these booming trends declined after 2020, which might be influenced by the time lag of self-constructed databases and the COVID-19 pandemic, especially

the following quarantine time and travel restrictions. In total, co-utilization has shown increasing trends in the past and might keep booming in the future.

The number distribution of publications related to the number of co-utilized facilities is shown in Figure 1(B). The number of co-utilized facilities increases by one unit, the number of publications might receive a tenfold decline approximately. In Figure 1(C), we recorded those highly frequent combinations and applied a linear fit to the distribution, contributing to describing the mechanism of facilities co-utilization. In the figure, almost every top choice shows great preference in geography that the facilities in the combinations might be in the same nation or region, for instance: both *PF* and *SPring-8* are Japanese facilities; *APS*, *NSLS-II*, and *ALS* are in the USA; *ESRF*, *Diamond*, and *PETRA* are in Europe. In total, more combinations involved might be a future trend and it is important to unveil the relationship between novel or common combinations and scientific breakthroughs and understand the impact of global technological co-utilization. In Figure 2, we could also observe the impacts of geographical factors in North America, Europe, and East Asia.

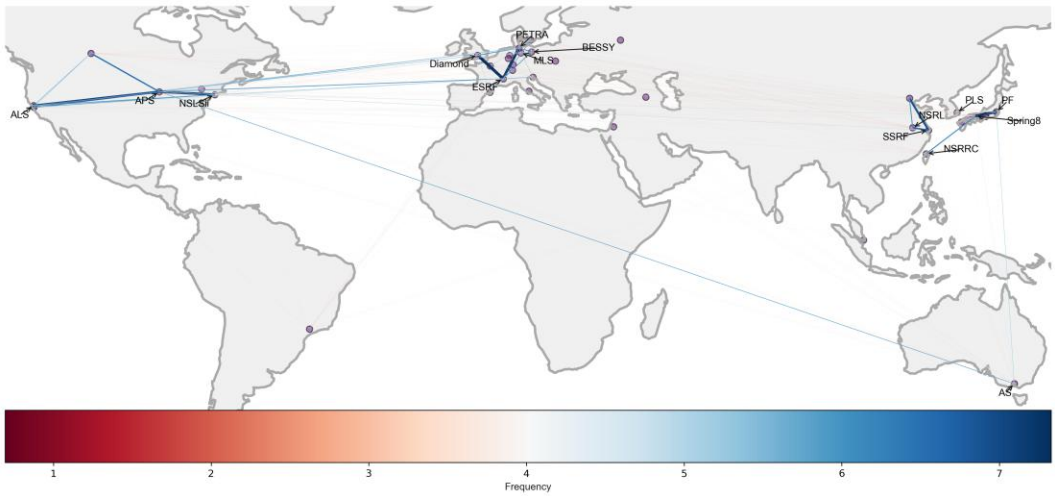


Figure 2. Global Distributions of Co-utilized Facilities.

We visualized the co-utilized relationships between global big science facilities and enclosed the names of the Top 15 facilities in productivity for better indication in Figure 2. The nodes in the figure represent big science facilities in our dataset and the links represent the frequency of co-utilizations between every two facilities with observations.

Potentially differences between Co-utilization and Singly utilization.

We adopted the Disruption Index (DI) as an indicator to measure the disruptive performance of scientific publications. DI was proposed by Funk and Owen-Smith (2017) and revised by Wu et al. (2019), and it has been widely used in scientometrics. Limited by the pages, we do not introduce this indicator in this progress work. In brief, if DI>0, indicating that the focal paper might bring a new orientation in knowledge system while DI<0, the focal paper might consolidate the current

knowledge system (Lin et al., 2023; Zhang et al., 2024). We mainly used the probability of disruption and considered disruptive publications are DI>=0.
Additionally, in the context of user-oriented big science facilities, there are two main research communities, External Scientists (Users, who visit the facility) and Internal Scientists (Staff, affiliated with the facilities), and the users are always in domination and the staff might be overlooked in the scientific publications since users might collaborate but will not co-author with them (D'Ippolito & Rüling, 2019; Söderström, 2023b). However, we demonstrate that co-utilization might bring more chances for internal scientists to co-author in user research.

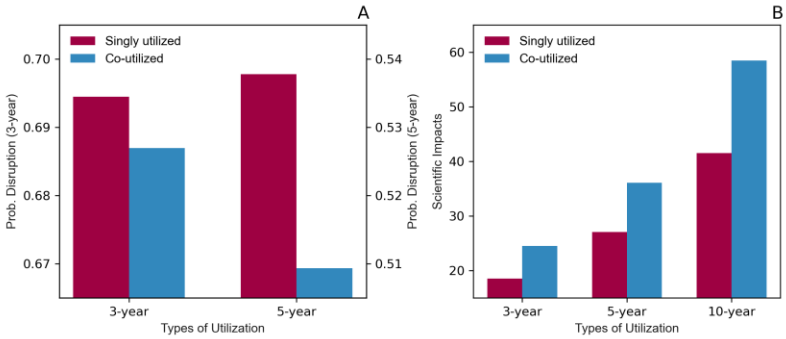


Figure 3. Performance Differences in Disruption and Scientific Impacts between Co-utilization and Singly Utilization.

In Figure 3, we mainly displayed the performance gaps between co-utilizations and single utilization by measuring the disruptive probability (A) and scientific impacts (B) of their supporting publications. Singly utilizations might produce more disruptive knowledge but receive fewer citations than co-utilizations since published in a 3-year, a 5-year, and a 10-year citation window.

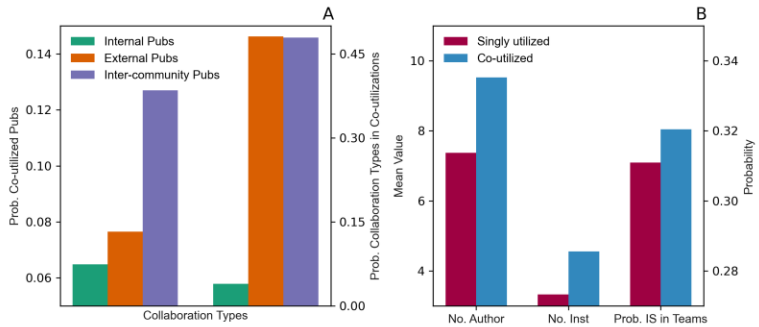


Figure 4. Differences between Co-utilizations and Single utilizations.

In Figure 4(A), we observe that above 12% of inter-community publications are supported by more than one facility (co-utilization) and the value is much higher than the ratio (above 7%) in External publications (authored by external scientists at all). Moreover, in the dataset of co-utilizations (23,046 papers are mentioned in Table 1), the ratios of inter-community publications and external publications are close, which also reveals that co-utilizations might provide more chances for staff participation. In Figure 4(B), we demonstrate that co-utilization might involve more

authors and institutions in collaboration and the probability of internal scientists participating in teams is also higher than single utilization, which further ensures that co-utilization might balance the community participation.

Conclusion and Future Works

This research in progress mainly reports a novel type of scientific collaboration based on a unique dataset of publications collected by us by crawling or exporting bibliography from SLSs' self-constructed databases. Future works could further explore the relationships between co-utilizations and scientific performance in the context of a resource-based view and the theory of S&T human capital. Moreover, we would also compare the main differences between facility co-utilization and inter-organizational collaboration in academia.

Acknowledgments

This work was supported by the National Social Science Fund Major Projects of China (Project No. 22&ZD127). We would like to thank Lingling Zhang, Yuhui Dong, and Honghong Li for their expertise and assistance with the basic knowledge of Large-scale Research Infrastructures and valuable comments from reviewers.

References

Bianco, W., Gerhart, D., & Nicolson-Crotty, S. (2017). Waypoints for Evaluating Big Science*. *Social Science Quarterly, 98*(4), 1144-1150. doi:10.1111/ssqu.12467

Börner, K., Silva, F. N., & Milojevic, S. (2021). Visualizing big science projects. *Nature Reviews Physics, 3*(11), 753-761. doi:10.1038/s42254-021-00374-7

D'Ippolito, B., & Rüling, C. C. (2019). Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications. *Research Policy, 48*(5), 1282-1296. doi:10.1016/j.respol.2019.01.011

Funk, R. J., & Owen-Smith, J. (2017). A Dynamic Network Measure of Technological Change. *Management Science, 63*(3), 791-817. doi:10.1287/mnsc.2015.2366

Hallonsten, O. (2014). How expensive is Big Science? Consequences of using simple publication counts in performance assessment of large scientific facilities. *Scientometrics, 100*(2), 483-496. doi:10.1007/s11192-014-1249-z

Heidler, R., & Hallonsten, O. (2015). Qualifying the performance evaluation of Big Science beyond productivity, impact and costs. *Scientometrics, 104*(1), 295-312. doi:10.1007/s11192-015-1577-7

Heinze, T., & Hallonsten, O. (2017). The reinvention of the SLAC National Accelerator Laboratory, 1992-2012. *History and Technology, 33*(3), 300-332. doi:10.1080/07341512.2018.1449711

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1-18. doi:10.1016/s0048-7333(96)00917-1

Lauto, G., & Valentin, F. (2013). How Large-Scale Research Facilities Connect to Global Research. *Review of Policy Research, 30*(4), 381-408. doi:10.1111/ropr.12027

Lin, Y., Frey, C. B., & Wu, L. (2023). Remote collaboration fuses fewer breakthrough ideas. *Nature, 623*(7989), 987-991. doi:10.1038/s41586-023-06767-1

Priem, J., Piwowar, H. A., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv, abs/2205.01833*.

Silva, F. S. V., Schulz, P. A., & Noyons, E. C. M. (2019). Co-authorship networks and research impact in large research facilities: benchmarking internal reports and bibliometric databases. *Scientometrics, 118*(1), 93-108. doi:10.1007/s11192-018-2967-4

Söderström, K. R. (2023a). Global reach, regional strength: Spatial patterns of a big science facility. *Journal of the Association for Information Science and Technology, 74*(9), 1140-1156. doi:10.1002/asi.24811

Söderström, K. R. (2023b). The structure and dynamics of instrument collaboration networks. *Scientometrics, 128*(6), 3581-3600. doi:10.1007/s11192-023-04658-w

Söderström, K. R., Åström, F., & Hallonsten, O. (2022). Generic instruments in a synchrotron radiation facility. *Quantitative Science Studies, 3*(2), 420-442. doi:10.1162/qss_a_00190

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378-382. doi:10.1038/s41586-019-0941-9

Zhang, M.-Z., Wang, T.-R., Lyu, P.-H., Chen, Q.-M., Li, Z.-X., & Ngai, E. W. T. (2024). Impact of gender composition of academic teams on disruptive output. *Journal of Informetrics, 18*(2), 101520. doi:https://doi.org/10.1016/j.joi.2024.101520

https://doi.org/10.51408/issi2025_138

Algorithmically Calculated Mentorship: The Netherlands Validation Study

Kathryn O. Weber-Boer[1], Carlos Areia[2], Tamarinde Haven[3]

[1]*k.weberboer@digital-science.com*,[2]*c.areia@digital-science.com*
Cornell University, Digital Science (USA)

[3]*T.L.Haven@tilburguniversity.edu*
Tilburg University (Nederlands)

Abstract

Many efforts to intervene in research practices, with the aim of promoting open science and research integrity, are based on intuitive speculation about what actions might be effective. Research culture involves the norms for registration, research, and publication, but also what responsibility is taken for training, and which behaviours are conventional in collaboration. Efforts to drive shifts in research culture often focus on awareness-raising activities, based on the assumption that a lack of knowledge or familiarity hinders practices of openness and integrity. These activities can be resource intensive, and participants may be self-selecting (where participation is voluntary). The hope is that awareness will spread organically into departments and disciplines. Testing the assumptions upon which these interventions are based provides data-driven evidence to support and strengthen these efforts.

One of the assumptions we are working to test is that open science and integrity practices are related to mentorship, or whether these practices are driven by other forces (e.g., career stage, national or institutional policies). In order to enhance the effectiveness of interventions, we seek to contribute to efforts to quantify the impact of mentorship on open science and research integrity practices. The research in progress presented here takes a first step in this quantification, by testing the foundation of a systematic approach to identifying mentor-mentee pairs. The ability to identify mentorship relationships at scale will enable the analysis of the relationship between mentor and mentee research practices, as well as allow for the assessment of other variables.

This work compares a manually curated dataset of candidates with PhDs awarded from 2021 and 2022 by four Dutch university medical centers and their supervisors (supervisory), to a dataset of pairs of researchers in which a mentor-mentee relationship was algorithmically determined (mentorship). All but one of the supervisory pairs were found in the mentorship dataset, and the strength of mentorship likelihood was largely high or very high. The mentorship dataset further includes informal mentors for the junior researchers. This lays the groundwork for a comparison of the research culture practices of supervisors and supervisees, compared to mentors from formal and informal relationships. This research so far demonstrates high confidence for algorithmically determined mentorship.

Introduction

The broader work of which this is a part aims to investigate the transmission of research culture between supervisors and supervisees. It is essential to be able to qualify and quantify the effect of research policy on research practice, to demonstrate the potential effects of incentives on open science practices. Without knowing whether there is an effect, we are limited in our ability to advocate for training programs, codes of conduct, or other efforts to enhance desirable research practices (Haven, 2025). Interventions in good research practice can be very resource

intensive, and a data-based assessment of the efficacy of these interventions would be valuable to the community.

Scholarly mentorship may play a vital role in shaping the careers of early-stage researchers. However, the impact of mentorship on research culture and practices is relatively under-explored. Correlation between mentorship, research integrity and open science practices remains unknown and there is a need for investigation to quantify the impact of mentorship and identify the factors that contribute to impactful relationships. At Digital Science [REF], we calculated billions of researcher to researcher relationships, including mentorship, however validation is required to test the accuracy and generalisability of this algorithm.

Therefore, the aim of this work is to establish the validity of a mentorship algorithm, by ensuring that manually curated supervisor-PhD pairs are identified in the resulting mentorship dataset, evaluating whether the strength of the relationship correlates with formal supervision, and assessing whether the mentorship dataset also provides likely candidates for informal mentors.

Mentorship is algorithmically determined by drawing upon evidence in publication and grant metadata for collaboration, combined with researcher-specific evidence of seniority. The algorithm produces a dataset of researcher pairs with numeric estimates of the closeness of the relationship and the degree and direction of seniority. The curated list of supervisor-PhD pairs was manually collected as part of a previous research project. This list is used to evaluate the accuracy of the mentorship dataset.

Methods

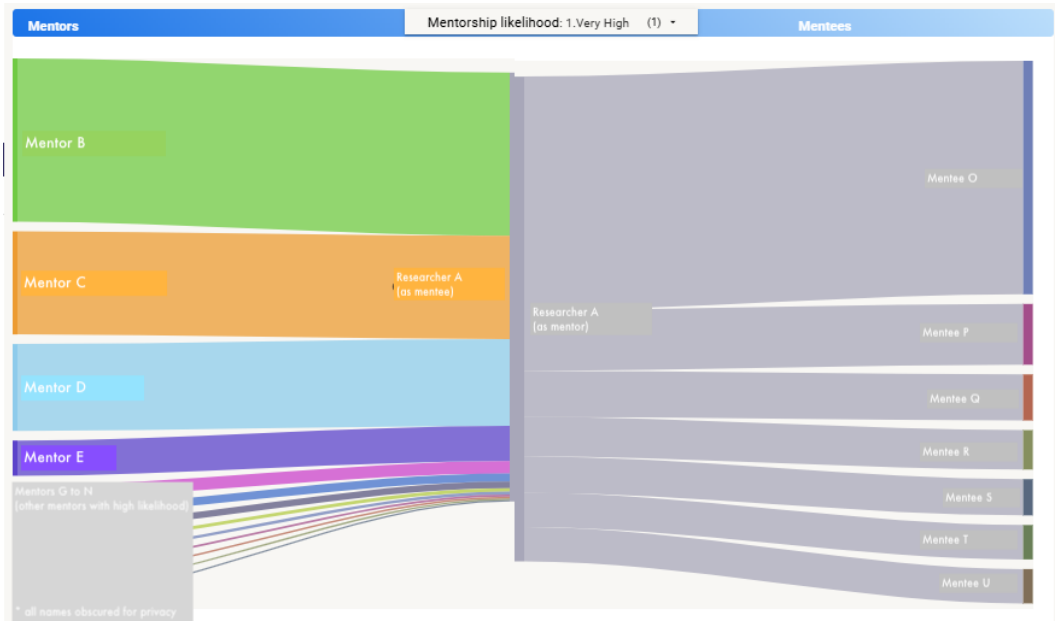
Curated PhD supervisor pairs

The curated PhD-supervisor pair dataset was curated as part of a process that developed new methods for quantifying role modelling of open science practices; the data are publicly available online. (https://github.com/tamarinde/ResponsibleSupervision/tree/main/Pilot-responsible-supervision). The data was manually collected based on a standardized protocol where researchers used PhD thesis metadata to systematically gather data about PhD candidate and main PhD supervisor (in Dutch: *promotor*). We uploaded the tables of PhD-supervisor pairs from four Dutch UMCs (Amsterdam, Groningen, Leiden, Maastricht) into Google BigQuery, and joined them into a final table. The data were cleaned for consistency. The resulting table consists of the PhD candidate and supervisor names, pair ID, and their publication DOIs, amongst other data (all publicly available in GitHub).

Mentorship Dataset

The mentorship database is composed of pairs of co-authors, with a calculated relationship strength and seniority estimation between two researchers: the mentor (Researcher) and the mentee (Co-Researcher). Both Dimensions and Altmetric data are used for this calculation, which considers:

1- Strength of the relationship, that includes data like the number of publications, citations and attention information (according to Altmetric data) of joint publications, grants and clinical trials; number of years sharing research, number of years in the same institution, and publication age.
2- Specific indicators of mentorship, including authorship position on the Researcher's first publications, and investigator roles in the Researcher's first grants.
3- Direction of the relation, using a seniorship score, where a higher score indicates the mentor in the Researcher-Co-Researcher pair.
Based on the above three points we calculated the strength and direction of the mentorship score, which we will refer to as the "Mentometer". A positive Mentometer score indicates that the Researcher is the mentor and the Co-Researcher is the mentee. We also used the Mentometer to calculate a categorical variable of the mentorship likelihood that ranged from "Very Low" to "Very High".



Dimensions data matching

To be able to match Dimensions researcher data to the correct PhD candidate and supervisors, we have followed these steps:

1. Grouped all DOIs available for each researcher
2. Extracted all authors names for each researcher publications
3. Tried to match all original tables PhDs/supervisors with the correct Dimensions researcher ID using the first 2 letters of the first name and the last 2 letters of the last name for the PhD candidates or last 3 letters of the last name for supervisors.
4. Ranked each researcher-Dimension author match automatically and only selected the top match
5. Two independent researchers (CA and KB) manually cross-checked the final PhDs and supervisors matching list, deleting abnormal matches when

multiple matches occurred, so there was only one researcher-Dimensions author match per PhD candidate and supervisor

6. Finally, the matched table includes the pair ID, the PhD candidate name, the supervisor name, subfield, and the thesis year from the curated PhD-supervisor pair dataset, and the supervisor and PhD candidate researcher identifiers from the Dimensions researchers dataset.

Using this linked dataset, we pulled the pairs of researchers from the Mentorship dataset and extracted the likelihood value of the mentor relationship of the supervisor-PhD candidate pairs. We then looked at other mentor candidates to establish whether there were stronger candidates identified in the mentorship algorithm.

One feature of the Dimensions researchers dataset is a tendency to privilege precision over recall. That is, whereas one researcher profile is highly unlikely to contain publications which are not authored by that researcher, it is not unexpected to find multiple profiles per researcher. We selected the strongest mentorship relationship pair, since there were a number of occasions on which multiple mentor-mentee pairs were found (representing the same PhD-supervisor pair). We also alerted the Dimensions support team of any duplicate researcher profiles found, for merging.

Results

This study included 213 distinct supervisors and 213 PhD candidates, all successfully matched to their respective Dimensions researcher IDs.

Table 1. Datasets and the number of supervisor and PhD names and pairs per set.

Dataset	*Supervisor Names*	*PhD Candidate Names*	*Pairs*
Manually curated	219	214	213
Matched Dimensions Researcher profiles	220	220	228
Matched in Mentorship dataset	214 (218 IDs)	213 (218 IDs)	212

Of the 213 PhD-supervisor pairs, 212 were found as pairs in the Mentorship dataset. Because of the additional Dimensions researcher profiles per researcher, there were more mentorship pairs than PhD-Supervisor pairs. Of the mentorship pairs, 188 were classified with a very high likelihood mentorship, and a further 11 had a high likelihood. The remainder of PhD-Supervisor pairs had likelihood of medium, low, or very low. One pair was not identified (Table 2).

Table 2. Mentorship likelihood and pairs matched from manual dataset.

Dataset	*Unique Pairs*
1. Very high	187
2. High	11
3. Medium	7
4. Low	4
5. Very Low	3
not identified	1

Discussion

This study aimed to validate an algorithmic calculation of researchers' mentorship relation. The mentorship score under validation was algorithmically determined by drawing upon evidence in publication and grant metadata for collaboration, combined with researcher-specific evidence of seniority. The algorithm produced a dataset of researcher pairs with numeric estimates of the closeness of the relationship and the degree and direction of seniority. This algorithmically determined mentorship dataset has been previously used by two of the authors (CA and KWB) to explore the transmission of open access publication practices. While the results were promising, suggesting a positive correlation between the open access publishing of the supervisor and the open access publishing rate of the supervisee, three questions required additional investigation: 1) did the relationships identified by the algorithm reflect real-life supervision, and 2) how does the influence of informal mentors on research and publication behavior compare to that of formal supervisors? The research presented in this paper addresses the first question.

Our results support the use of our algorithm in similar populations, as the majority of the manually curated supervisions were identified by our algorithm as having "Very High" likelihood of mentorship. To our knowledge, this is the first study to validate an algorithmic-based mentorship relationship calculation amongst researchers. This validated algorithm will open the door to future exploration of the effect of these relationships on other research practices.

These results also increase our confidence in calculating and using our mentorship algorithm at scale within similar fields included in this dataset, often including millions of mentor-mentee pairs. The authors plan to use these manually curated supervisor relationships and algorithmically determined mentorship relationships to evaluate the role mentorship plays in the transmission of research culture, including open science practices such as ethical approval statements, authorship contribution statements, and data and code sharing. This research also serves as the foundation for other types of analysis, such as geographical mobility and impact related to mentorship, amongst others.

Limitations

In the manually curated dataset, we identified a number of PhD-supervisor pairs where the names of either the supervisor or the PhD candidate varied (e.g., middle initial vs. full middle name). This is a valuable data artefact, as it demonstrates the

limitations of manual curation. Conversely, a major limit of algorithmic identification is the inability to distinguish formal mentorship from informal with certainty.

This work is focused on the biomedical field (specifically researcher pairs from medical centers in the Netherlands). There will be fields for which this approach is less well-suited. Future work will explore these limitations.

Despite encouraging results, we acknowledge that the results of this study may only be generalizable within biomedical and clinical fields, and other validation is required in other fields. For example, we foresee our algorithm performance to be affected in fields where authorship behaviours are different than in medical fields (for example, in mathematics where authorship is usually alphabetical, or the humanities where single authorship is more common).

Acknowledgments

The authors acknowledge the valuable contribution of Susan Abunijla and Nicole Hildebrand, who helped collect, curate, and analyse the manually curated dataset (Haven et al., 2023).

References

Haven, T. (2025). It takes two flints to start a fire: A focus group study into PhD supervision for responsible research. Accountability in Research, 1–24. https://doi.org/10.1080/08989621.2025.2457584

Haven, T.L., Abunijela, S., Hildebrand, N. (2023). Biomedical supervisors' role modeling of open science practices. *eLife*, 12:e83484. https://doi.org/10.7554/eLife.83484

Weber-Boer, K., Areia, C., and Taylor, M. (2024). *Is openness heritable: the transmission of integrity from mentor to mentee*. World Conference on Research Integrity. 3 June 2024.

https://doi.org/10.51408/issi2025_143

Application of Molecular Docking Technology in Drug Discovery Based on Bibliometric and Patent Analysis

Zhou Haichen[1], Jorge Gulín-González[2], Chen Yunwei[3]

[1]*zhouhc@clas.ac.cn*, [3]*chenyw@clas.ac.cn*
Scientometrics & Evaluation Research Center (SERC) of National Science Library (Chengdu), Chinese Academy of Sciences (People's Republic of China)

[2]*gulinj@uci.cu*
Centro de Estudios de Matemática Computacional (CEMC). Universidad de las Ciencias Informáticas (UCI), La Habana (Cuba)

Abstract

In the field of molecular modeling, molecular docking (MD) is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using scoring functions. MD is frequently used to predict the binding orientation of small molecule drugs candidates to their protein targets in order to in turn predict the affinity and activity of the small molecules. MD plays an important role in the rational design of drugs. On the other hand, patents are a significant output-based indicator of innovation, analyse countries and organizations' technological capabilities, analyse relationship between polices and technological innovation. In this context, it is interesting to investigate the dynamics of the behaviour of granted patents and scientific articles related to them in the topic of MD, with the development of new drugs. In this paper we present a comprehensive assessment about the application of MD technology in drug discovery through bibliometric and patent analysis, revealing research trends, technological hotspots, key participants and their collaboration networks, as well as the connection between academic research and practical applications. The study covers the period 1979-2024. We use a keyword co-occurrence network to analyze high-frequency keywords in publications and identify research hotspots and we apply BertTopic to extract research topics and their evolution over time. Also, it is our objective reveal development patterns and critical milestones in molecular docking technology for drug discovery. Our research provides research hotspot references for academia and strategic insights for industry stakeholders and promotes collaborative innovation between academic research and industrial practice. In future works we will present the analysis of a study case related to the development of drugs for the treatment of COVID-19.

Introduction

The identification of drug candidates is one of the most arduous stages in the design of new drugs (Schnecke & Boström, 2006). Molecular docking (MD) is a technique which examines the conformation and orientation of molecules, mainly ligands, into the binding site of a protein target. Searching algorithms generate likely poses, which are ranked by scoring functions (Liu et al., 2018). To generate a receptor (protein)-ligand structure *in silico* two steps are followed: (i) Docking per se entails conformational and orientational sampling of the ligand within constraints of the receptor binding site and, (ii) Scoring function selects the best pose for a given molecule and rank orders ligands. At present, several software are available to carry out MD, among them: AutoDock (Morris et al., 1998), AutoDock Vina (Trott & Olson, 2010), DockThor (DeMagalhães, et al., 2014), FlexX (Rarey et al, 1996) and

GOLD (Verdonk, M.L,2003). The number of scientific publications related to MD has been increasing significantly in the last 25 years.

On the other hand, patents are as a significant output-based indicator to measure innovation, analyse countries and organizations' technological capabilities, analyse relationship between polices and technological innovation (Banerjee et al., 2000). Interesting studies on the relationship between patents and the economic development of a country, particularly in the field of biotechnology, have been published recently (Qiang et al. 2019). The analysis of the number of patents can give us an idea of the impact of theoretical and computational techniques such as DM on the development of new drugs. These results can be complemented with an analysis of the number of publications per year, their trend and relation with the number of patents, as well as by the analysis of other scientometric parameters.

In this paper we present a comprehensive assessment about the application of MD technology in drug discovery through bibliometric and patent analysis, revealing research trends, technological hotspots, key participants and their collaboration networks, as well as the connection between academic research and practical applications. This research provides research hotspot references for academia and strategic insights for industry stakeholders and promotes collaborative innovation between academic research and industrial practice.

Methods

The selected scientific papers encompass science citation index expanded (SCIE) articles and conference proceedings citation index–science (CPCI-S) proceedings papers from 1979 to 2024, sourced from the Web of Science Core Collection with the search strategy as: TS=("molecular docking" OR "docking") AND TS=("drug* discover*" OR "ligand*" OR "drug* design*") AND PY=1979-2024. We collected 38240 results in the studied period. For this study, only research papers (articles and proceedings papers) were selected, since they are directly related to the study and development of new drugs. The selection of SCIE as the database for this first part of the study was based on the fact that it is the most internationally recognized database and since molecular docking is a novel and high-impact topic, it is to be expected that a significant percentage of the main research papers will be published in journals from this collection. Patent data information related to molecular docking from the PatSnap Analytics database (Alkhazaleh R & Mykoniatis, 2024), with the search strategy as: Keywords=(("molecular docking" OR "docking")" AND "("drug design" OR "drug discovery" OR "ligand")") AND Publication Date="1979/01/01-2024/12/31". The retrieval date is January 2025. The PatSnap's advantage lies in its coverage of 172 patent offices worldwide, containing over 1.96 billion patent records. With daily updates, it ensures real-time access to comprehensive and up-to-date global patent data, supporting precise and efficient intellectual property research.

Results and Discussion

Overall tendency

In Figure 1 we present the annual publication volume related to MD topic over the studied period 1979-2024. A sustained increase in the number of papers can be observed, this increase is even more significant from 2020, which can be related to the health emergency resulting from the COVID-19 pandemic. As an example of that, the most cited docking software AutoDock is used by the FightAIDS@Home and OpenPandemics - COVID-19 projects run at World Community Grid, to explore for antivirals in the treatment of HIV/AIDS and COVID-19. Also, AutoDock has contributed to the design of relevant drugs, including HIV1 integrase inhibitors (Goodsell et al., 2021). Figure 1 also shows the number of patents per year linked to MD between 1979 and 2024. According to the behavior of the numbers of papers and patents, graph of Figure 1 can be divided in three regions: (i) The first goes up to 1999 and both the number of papers and patent applications are very similar and relatively low, which corresponds to the incipient state of research on the subject and the limited computational capabilities in that period. Also, it is expected that at the beginning prevail research of a theoretical nature, (ii) The second period (between 2000 and 2019), in which there is initially an increase in papers and patent application, then, quickly, in the case of patent application, a plateau (2000-2012) is reached. In our opinion, the initial increase corresponds to a higher degree of maturity of the molecular docking technique; the plateau can be explained from the accumulation of knowledge in the previous period and the non-existence of exceptional pandemic episodes at these years. In the interval 2013-2020 both parameters have practically the same behavior, (iii) After 2020, both papers and patent application increase significantly, which is directly related to the appearance of the COVID-19 pandemic.

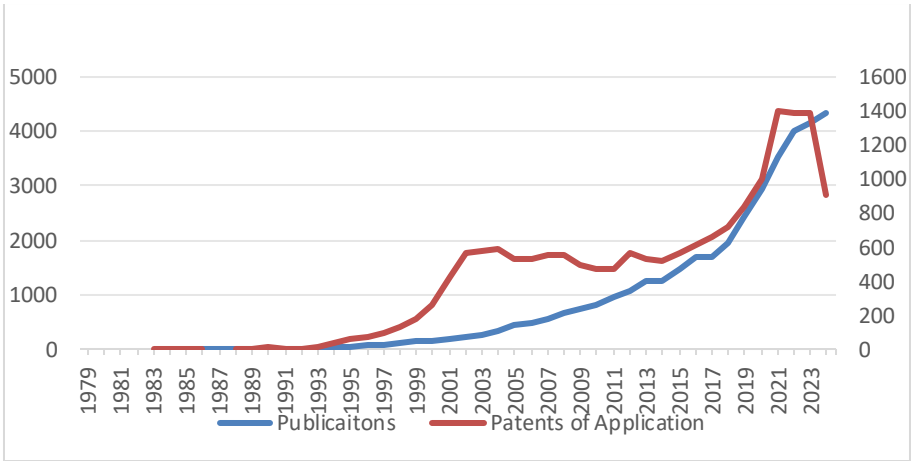


Figure 1. Annual publications and patents related to MD topic in the studied period 1979-2024.

Note: The patent application data for 2023 and 2024 has not yet been fully disclosed.

Other interesting aspect is to study funding availability. According to the dataset of papers (1979 - 2024), 20,502 papers were funded by 973 agencies and institutions from 76 countries/regions. It is a 53.6% of all the published papers. The annual trend of support of five representative funding agencies and institutions are shown in Figure 2. In the last 15 years there has been a sustained growth in funding for this type of research, especially after the COVID-19 pandemic, with funding from the National Natural Science Foundation of China (NNSCF) standing out in this last period.

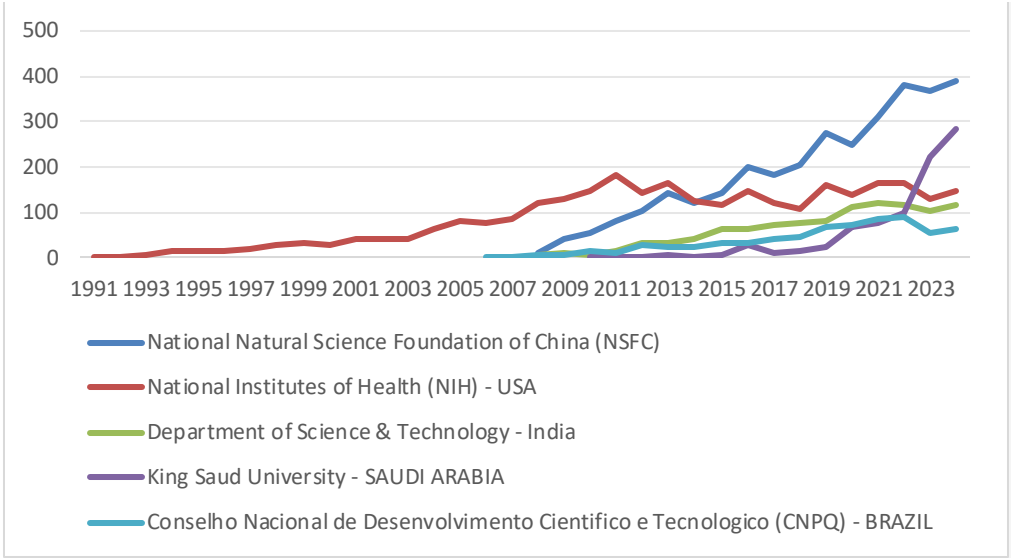


Figure 2. MD publication counts funded for five representative agencies and institutions in the studied period.

Research Hotspot Analysis

In Figure 3 we present a keyword co-occurrence network to reveal the research hotspot in MD. Six hotspots identified are as followed:

1. Drug Design and Discovery (Red Cluster)

This cluster highlights the application of molecular docking in identifying and optimizing potential drug candidates. Keywords such as "virtual screening", "drug design", "prediction", and "binding-affinity" underscore the reliance on computational methods for screening and improving ligand-target interactions. Terms like "flexibility", "force-field", and "genetic algorithm" point to methodological advancements to enhance docking precision and efficiency.

2. Structural and Biological Validation of Molecular Docking (Blue Cluster)

This cluster integrates structural biology, experimental validation, and biological activity evaluation to support molecular docking predictions. Core keywords include "crystal-structure", "x-ray", "fluorescence spectroscopy", and "protein-binding", reflecting the reliance on high-resolution structural data to model and verify molecular interactions. Simultaneously, terms such as "antibacterial", "antioxidant

activity", and "cytotoxicity" emphasize the application of docking techniques in identifying compounds with specific therapeutic properties. The presence of "in-vitro" further indicates the integration of computational predictions with experimental validation to confirm biological relevance.

3. Computational Methods and Algorithms (Yellow Cluster)
This cluster represents the methodological and algorithmic innovations in molecular docking studies. Keywords like "3D-QSAR", "pharmacophore", and "homology model" highlight the use of advanced modeling techniques to predict and optimize molecular interactions. The inclusion of "protein-coupled receptor", "GPCR", and "kinase inhibitors" demonstrates the diverse range of target molecules addressed by these computational methods.

4. Biological Activity and Therapeutic Applications (Green Cluster)
This cluster explores the application of docking in discovering and evaluating compounds for their biological activity and therapeutic potential. Terms such as "cancer", "resistance", "drug discovery", and "biological evaluation" reflect the use of docking for identifying bioactive molecules in disease-related contexts. Keywords like "activation", "target", and "expression" emphasize the subsequent experimental validation of docking results through pathway analysis and functional assays.

5. Neurological and Viral Diseases (Yellow-Green Cluster)
This cluster bridges molecular docking with research on neurological and viral diseases, addressing global health challenges. Key terms such as "Alzheimer's disease", "oxidative stress", "virus", and "SARS-CoV-2" suggest a focus on identifying therapeutic molecules for these diseases. The presence of "network pharmacology" and "pathogenesis" indicates a systems-level approach to understanding disease mechanisms and therapeutic interventions.

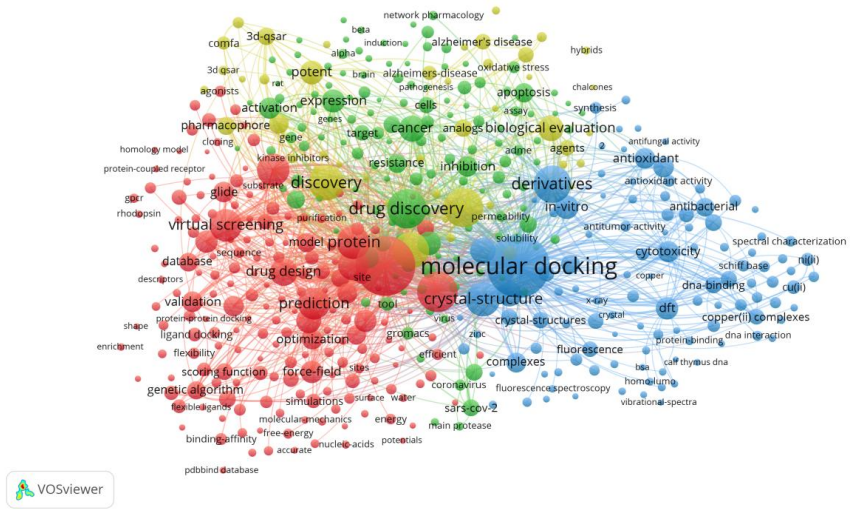


Figure 3. Keyword co-occurrence network in MD (Occurrences ≥ 100).

National/Regional Collaboration Networks

In Figure 4 we present the national collaboration in MD. A short description of these regions regarding to the MD techniques is shown in the following.

1. North America and East Asia as Pioneering Regions in MD Research

North America, led by the United States, and East Asia, dominated by China, emerge as the most influential regions in the global molecular docking research network. The United States maintains its leadership position through pioneering computational methodologies and consistent funding for drug discovery initiatives, while China has rapidly ascended in this domain, leveraging its strategic investment in bioinformatics and increasing collaboration with both developing and developed nations.

2. Europe's Collaborative Network in Multidisciplinary Research

European countries, notably Germany, Italy, and France, form a dense and interlinked cluster, underscoring the region's collaborative research culture. This interconnectedness stems from European Union funding initiatives, such as Horizon Europe, which promote cross-border partnerships and multidisciplinary projects. The European network's focus likely extends to fundamental research and the development of innovative molecular docking algorithms, emphasizing their application in diverse fields such as cancer therapeutics and personalized medicine. The prevalence of collaboration within this cluster highlights the synergistic potential of smaller countries working collectively with leading research nations.

3. South Asia and the Middle East as Emerging Contributors

The emergence of South Asia and the Middle East, represented prominently by India and Saudi Arabia, reflects the growing contributions of these regions to molecular docking research. These countries prioritize computational approaches to tackle region-specific health issues, such as neglected tropical diseases and antimicrobial resistance. India, in particular, has become a regional hub for bioinformatics research, leveraging its strong pharmaceutical industry and a rapidly expanding academic infrastructure. Saudi Arabia's position suggests strategic investments in life sciences, likely aimed at diversifying its research portfolio and fostering international collaborations in computational drug discovery.

4. Cross-Cluster Collaboration Driven by Global Health Challenges

The strong interconnectivity across clusters signifies a global effort to address pressing biomedical challenges through molecular docking. Two major themes emerge from these collaborations: the discovery of antiviral agents, accelerated by the COVID-19 pandemic, and the development of targeted therapies for cancer and other chronic diseases. These thematic focuses necessitate partnerships that bridge technological expertise, such as that found in North America and East Asia, with diverse research perspectives from Europe, South Asia, and the Middle East. The network visualization thus underscores the centrality of molecular docking as a

unifying research domain, fostering innovation through cross-border scientific exchange.

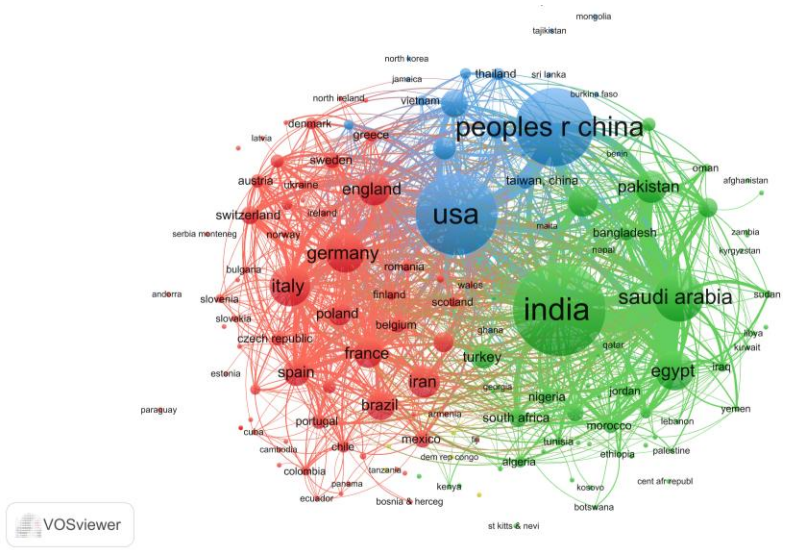


Figure 4. National/ regional collaboration in MD.

Institutions with the most patent applications

In Table 1 we present the Top 20 current assignee of patent in MD. Academic and research-focused organizations account for a significant share of patent filings, with The Regents of the University of California leading with 381 patents. Universities such as Sanskriti University, Harvard College, and Stanford University follow suit. These institutions dominate the rankings, reflecting their emphasis on foundational research and innovation. Novartis AG (119 patents), Bristol-Myers Squibb (92 patents), Genentech, Inc. (89 patents), and Allergan, Inc. (75 patents) are prominent pharmaceutical companies. Their focus is on translating molecular docking innovations into clinical applications. Research hospitals and non-profit institutions also feature prominently in this domain. Notable examples include:Dana-Farber Cancer Institute, Inc. (109 patents), City of Hope (84 patents), The General Hospital Corporation (70 patents).

Table 1. Top 20 current assignee of patent in MD in the studied period 1979-2024.

Current Assignee	*Patent Count*
THE REGENTS OF THE UNIVERSITY OF CALIFORNIA	381
SANSKRITI UNIVERSITY	246
PRESIDENT AND FELLOWS OF HARVARD COLLEGE	137
THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY	131
NOVARTIS AG	119
DANA-FARBER CANCER INSTITUTE, INC.	109
MASSACHUSETTS INSTITUTE OF TECHNOLOGY	109
IMMUNOMEDICS, INC.	104
BOARD OF REGENTS, THE UNIVERSITY OF TEXAS SYSTEM	97
BRISTOL-MYERS SQUIBB COMPANY	92
YALE UNIVERSITY	91
THE SCRIPPS RESEARCH INSTITUTE	90
GENENTECH, INC.	89
CITY OF HOPE	84
THE BROAD INSTITUTE, INC.	79
ALLERGAN, INC.	75
THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK	74
THE GENERAL HOSPITAL CORPORATION	70
VANDERBILT UNIVERSITY	70
DUKE UNIVERSITY	67

Conclusions

In this paper we present a comprehensive assessment about the application of MD technology in drug discovery through bibliometric and patent analysis, revealing research trends, technological hotspots, key participants and their collaboration networks. We found an increasing trend in the number of papers and patents in the field of MD in the studied period. Furthermore, a relationship has been found between the number of papers and patents. Also, a preliminary analysis of the funding of agencies and institutions for the support of research in MD was carried out. In our study, main countries and institutions with patents in the drug design have been identified. Our results can help to better understand the dynamic relationship between scientific work expressed in papers and the development of new drugs based on the number of patents granted. Future developments will focus on quantifying the relationships found and studying these relationships in detail for a case study related to the development of drugs for the treatment of COVID-19.

Acknowledgments

This work was supported by the key project of innovation fund from National Science Library
(Chengdu), Chinese Academy of Sciences (E3Z0000902).

References

Alkhazaleh, R., & Mykoniatis, K. (2024). Unveiling predictors influencing patent licensing: Analyzing patent scope in robotics and automation. *World Patent Information*, 77, 102276.

Banerjee, P., Gupta, B., & Garg, K. (2000). Patent statistics as indicators of competition an analysis of patenting in biotechnology. *Scientometrics*, 47(1), 95-116.

De Magalhães, C. S., Almeida, D. M., Barbosa, H. J. C., & Dardenne, L. E. (2014). A dynamic niching genetic algorithm strategy for docking highly flexible ligands. *Information Sciences*, 289, 206-224.

Goodsell, D. S., Sanner, M. F., Olson, A. J., & Forli, S. (2021). The AutoDock suite at 30. *Protein Science*, 30(1), 31-43.

Liu, Z., Liu, Y., Zeng, G., Shao, B., Chen, M., Li, Z., ... & Zhong, H. (2018). Application of molecular docking for the degradation of organic pollutants in the environmental remediation: A review. *Chemosphere*, 203, 139-150.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry*, 19(14), 1639-1662.

Qiang L., Chen Y., Li X., Qi L., J. Gulín-González, Zhang Z (2019). A citation iteration method for publications and scientists' evaluation. (2021). *Data Science and Informetrics*, 1(2).

Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology*, 261(3), 470-489.

Schames, J. R., Henchman, R. H., Siegel, J. S., Sotriffer, C. A., Ni, H., & McCammon, J. A. (2004). Discovery of a novel binding trench in HIV integrase. *Journal of medicinal chemistry*, 47(8), 1879-1881.

Schnecke, V., & Boström, J. (2006). Computational chemistry-driven decision making in lead generation. *Drug discovery today*, 11(1-2), 43-50.

Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2), 455-461.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein−ligand docking using GOLD. Proteins: *Structure, Function, and Bioinformatics*, 52(4), 609-623.

https://doi.org/10.51408/issi2025_123

Automatic Literature Review Generation by Integrating Large and Small Models

Xiaofei Li[1], Guo Chen[2]

[1]358246618@qq.com,[2]delphi1987@qq.com
School of Economics & Management, Nanjing University of Science & Technology, Nanjing (China)

Abstract

This study proposes an innovative method for the automatic literature review generation, specifically designed for small-scale literature corpus in niche domains. First, the proposed method enhances the BERTopic with LLM to extract macro-level topic information. It then leverages LLM-based instruction fine-tuning to identify micro-level move structures. Finally, the method employs systematic LLM fine-tuning and a template-based generation strategy to automatically generate review that are thematically clear and logically coherent. Experimental results demonstrate that this method generates high-quality, well-structured review texts with clear topics when applied to small-scale citation analysis corpus, offering a new reference and practical example for automatic literature review generation in specialized fields.

Introduction

The development of automatic literature review generation techniques has significantly enhanced researchers' ability to efficiently acquire knowledge by synthesizing concise review texts from thematically related studies (Asmussen & Møller 2019). Current approaches to automatic literature review generation can be broadly classified into two categories: (1) extractive and generative methods based on small-scale deep learning models and (2) natural language generation methods leveraging LLMs. The first category integrates extractive and generative techniques, exemplified by the method proposed by Vaishali et al.(Vaishali et al., 2024), which employs an improved TextRank algorithm to extract key sentences from multiple documents, followed by a Seq2Seq model for review generation. While these methods are effective, they heavily depend on large-scale corpora and often struggle with maintaining consistency between the generated content and the original text. The second category, represented by retrieval-augmented generation approaches such as the one introduced by Han et al. (Han et al., 2024), incorporates relevant literature as an external knowledge source to enhance LLM-based review generation. Although these methods achieve superior fluency and logical coherence, they remain constrained by the inherent limitations of LLMs, including restricted context window sizes, outdated knowledge representations, and susceptibility to generating "hallucinated" information (Wang et al., 2024). Additionally, existing methods struggle to balance accuracy and comprehensiveness when processing small-scale literature corpus that are continuously updated in niche domains. To address these challenges, this paper proposes a novel hybrid framework that integrates small models with LLMs for automatic scientific review generation. The proposed approach leverages LLMs' strengths in language understanding and knowledge

synthesis while incorporating topic model to uncover latent topics, enabling high-precision review tailored to small-scale niche literature corpus.

Methodology

The proposed automatic review generation framework consists of three steps (as shown in Figure 1). First, a BERTopic model enhanced by LLM is used to identify the macro-level topic component of the document set. At the same time, the LLM is instruction-tuned to generate the micro-level move component of the documents. Finally, the LLM integrate two components based on a predefined template to generate a review. Each step is detailed as follows.

LLM-Enhanced BERTopic Zero-Shot Topic Modeling

Topic models are effective tool for extracting topic information from document set, providing a macro-level perspective for literature review. However, both traditional topic models such as LDA, Top2Vec, and BERTopic and LLMs face challenges when applied to the small-scale niche literature corpus. These challenges include sparse topic distributions due to the limited number of documents and high semantic overlap, which makes it difficult to distinguish topics. Moreover, these models struggle to capture deep topic relationships between documents, leading to unclear or even distorted topic results.

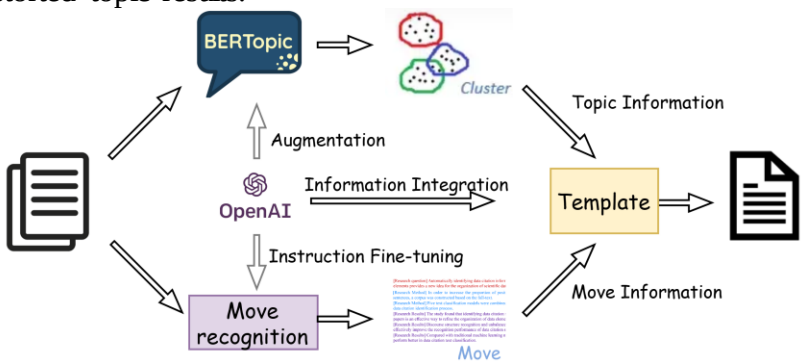


Figure 1. Framework of the Automatic Literature Review Generation Method.

To address these issues, this study proposes an LLM-enhanced zero-shot BERTopic modeling approach. This method integrates LLMs into three stages of topic modeling: enhancing document topic representation in the text embedding phase, assisting topic identification in the modeling phase, and refining topic distribution and representation in the post-modeling phase. This approach improves the overall performance of topic model.

In the text embedding phase, traditional word embedding models often fail to adequately capture intricate topic relationships between documents when processing small-scale literature collections in specialized domains. This limitation subsequently undermines the performance of topic model. To address this issue, this study leverages LLM to enhance document topic representation (as illustrated in Figure 2). The proposed methodology consists of two key stages: First, LLMs are utilized to generate high-quality topical phrases, tags, and descriptions from raw

document content, thereby extracting critical topic information. Second, distinct embedding representations are created for both the topic information and the original documents. These representations are subsequently fused via vector concatenation, thus creating a unified document embedding that highlights thematic features. This method improves the effectiveness of topic model by integrating topic-focused information with the original semantic content.

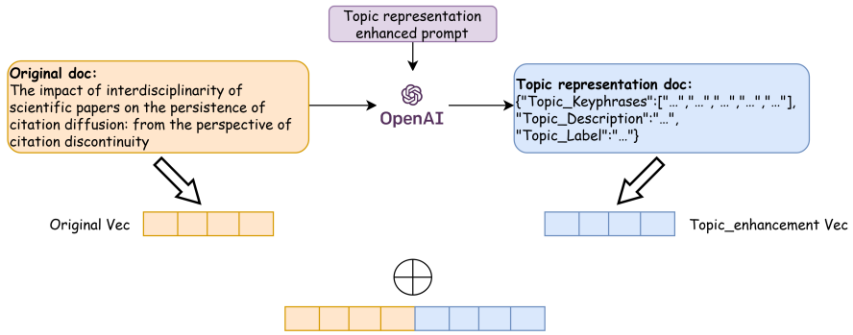


Figure 2. Method for Enhancing document Topic Representation.

In the topic identification phase, this study departs from the conventional approach of mining topics from scratch and instead leverages LLM-generated prior knowledge to guide BERTopic's zero-shot topic modeling (as illustrated in Figure 3). Specifically, LLMs are employed to extract salient topics from the document set, which are subsequently used as zero-shot topics for BERTopic. Following this, we calculate the similarity between each document and the zero-shot topics, then apply hierarchical processing based on similarity threshold: documents with similarity scores exceeding the threshold are directly assigned to their corresponding topics, while the remaining documents undergo further topic identification via BERTopic. This layered strategy effectively integrates the prior knowledge provided by LLMs with the adaptability of BERTopic, thereby improving the accuracy of topic recognition while ensuring broad coverage of topic distribution.

During the experiments, we observed that BERTopic's results exhibited loose topic representations and ambiguous boundaries. To address this issue, this study integrates LLMs in the post-modeling phase for refining topic representations and adjusting distributions. First, we utilize LLM to generate semantically compact and coherent topic labels based on the original outputs, thereby replacing BERTopic's native topic representations. Subsequently, by leveraging LLMs' zero-shot classification capability, we reassign the topic affiliations of boundary documents using the optimized topic labels as classification criteria. This two-step optimization strategy enhances both the accuracy and consistency of the final results.

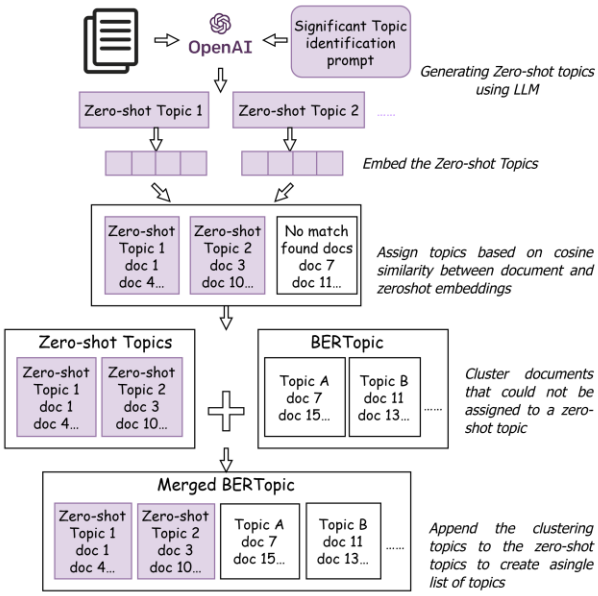


Figure 3. LLM-Guided BERTopic Zero-Shot Topic Modeling.

By integrating LLM into the entire workflow of BERTopic, our method achieves that the generated topics are not only semantically coherent and compact but also robust against noise and ambiguity in small-scale corpus.

Instruction fine-tuning LLM for Move Recognition

Move recognition effectively deconstructs sentence-level knowledge units in scientific literature, providing structured knowledge—such as research problems, methods, and results—that is essential for comprehensive review.

This study proposes an move recognition method that combines In-Context Learning (Agarwal et al., 2024) and Chain-of-Thought (Wei et al., 2022) LLMs prompting techniques, leveraging instruction-tuned LLM to accurately extract three types of knowledge units from abstracts: research problems, methods, and results. As illustrated in Figure 4, the prompt design for this method consists of four key modules: role setting and task description: guides the model to define its role and construct tasks based on instructions; Chain-of-Though: offers guided reasoning steps to help the model establish a clear logical chain during move recognition; In-Context Learning: provides examples of move recognition; and input integration: presents abstracts of scientific literature as input.

To comprehensively evaluate the performance of LLM in move recognition, we developed a systematic validation framework. The evaluation dataset is derived from our research group's previous move recognition projects, which include high-quality human-annotated data (Chen & Xu, 2019). Recognizing that human-annotated moves and LLM-generated moves may differ in wording but maintaining semantic equivalence, we introduced BERTScore(Zhang et al. 2020), a deep semantic matching metric, to effectively assess the semantic consistency of LLM-generated rhetorical moves.

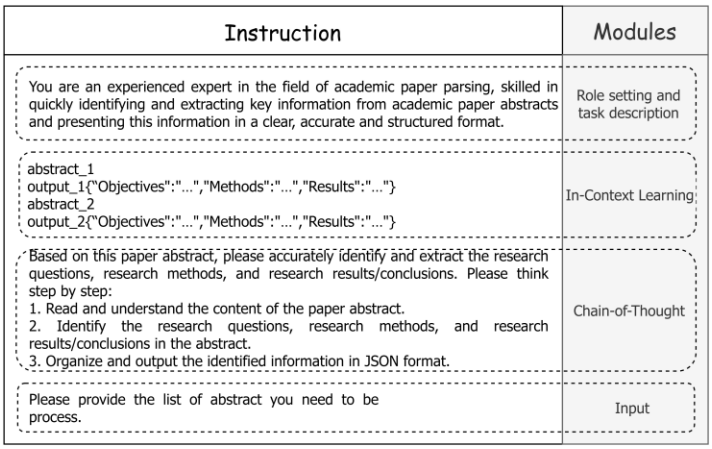


Figure 4. Move Recognition Prompt Design.

The comparative experimental results, summarized in Table 1, demonstrate the superior performance of instruction-tuned LLM in the move recognition task. Incorporating advanced prompt engineering strategies, the LLM-generated functional sentences exhibit significantly higher semantic consistency with human-annotated sentences. This enhanced performance establishes a reliable technical foundation for the task of automatic review generation.

Table 1. Comparative Results of Move Recognition.

Model	*Precision*	*Recall*	*F1-score*
SVM	0.834	0.834	0.834
CNN	0.839	0.838	0.839
Bi-LSTM	0.846	0.845	0.846
LLM	0.848	0.865	**0.853**

Template-Based Automatic Literature Review Generation with LLM

Through topic modeling and move recognition, we extracted both the macro-level topics and the micro-level move structures from the literature corpus. Subsequently, it is necessary to further investigate the internal connections between papers and construct a concise yet in-depth review. LLMs possess robust text generation and semantic integration capabilities. With instruction tuning, they can be effectively customized for literature review tasks. To enhance the effectiveness of LLMs in analysing topic connections and integrating moves, this study proposes a "Topic-Move" review template (see Figure 5) to standardize input data. Based on this, the automated review generation process consists of two stages. First, during the preprocessing stage, we organize the results of topic modeling and move recognition into a standardized and hierarchical format to ensure structured input. Then, in the generation stage, we employ chain-of-thought prompting combined with a modular generation strategy, completing the review in three steps: (1) feeding text segments with the same move under the same topic into the LLM to generate a move-level summary, (2) aggregating all move-level summaries under the same topic to produce

a topic-level summary, and (3) synthesizing all topic-level summaries to generate the complete review text.

In summary, the proposed automatic literature review method combines the strengths of LLMs and small models while overcoming their respective limitations. First, we leveraged the text comprehension and generation capabilities of LLMs to enhance topic representation and identify moves, thereby compensating for the semantic understanding shortcomings of small models when processing niche literature corpora. Second, the computational efficiency of small models is utilized for topic modeling, structuring raw literature into topic-level and move-level knowledge units — this dual approach reduces computational burdens on LLMs, thereby suppresses their hallucination tendencies. Furthermore, the framework adopts a phased generation architecture (move-topic-full-text) with modular strategies, effectively circumventing the context window constraints of LLMs.

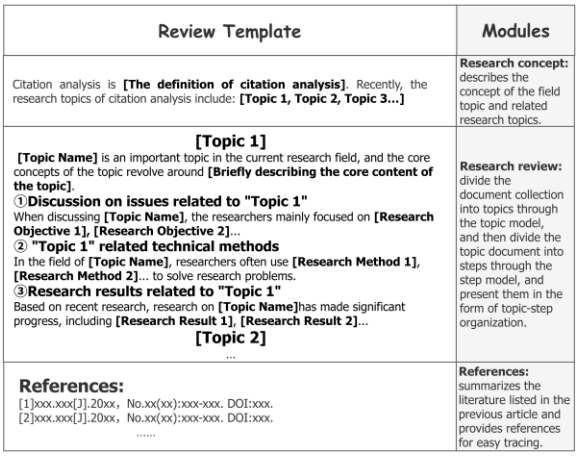


Figure 5. Template for Automatic Literature Review Generation.

Experiment – A Case Study in Citation Analysis

To evaluate the effectiveness of the proposed method, this study selected 24 papers in the field of "citation analysis" published in SSCI and CSSCI journals between September and December 2024 as experimental samples.

First, we utilized LLMs to enhance the topic representation of the original documents. (All LLMs used in this study were accessed via the OpenAI GPT-4 API). By inputting the titles, keywords, and abstracts of the papers, the LLM generated topical phrases, tags, and detailed descriptions. Subsequently, the model conducted a preliminary identification of significant topics within the dataset, recognizing three prominent topics (see Table 2). These identified topics were then employed as prior knowledge for zero-shot topic modeling using BERTopic. The preliminary modeling results (Figure 6) revealed one outlier topic, one zero-shot topic, and three topics derived through BERTopic.

Table 2. Significant Topics and Representations Identified by LLM (Partial Display).

Topic	*Topic Words*
Topic 0	['Impact', 'Measurement', 'Performance', 'Influence', 'Evaluation', …]
Topic 1	['Advanced Methods', 'Analytics', 'Novel Techniques', …]
Topic 2	['Data Management', 'Open Citation Data', 'Information Retrieval', …]

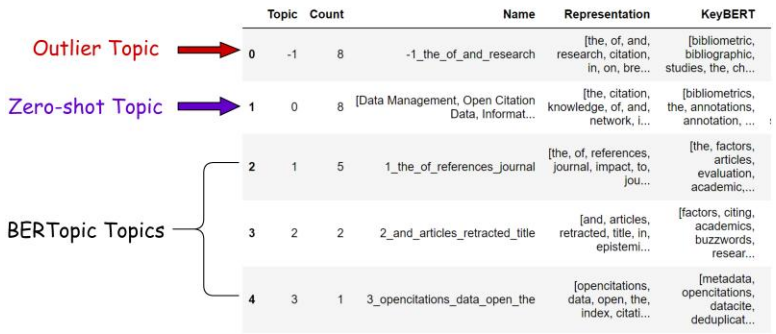


Figure 6. Preliminary Topic Modeling Results.

Following this, the LLM was employed to refine the topic representations and distributions. Specifically, the experiment integrated the topic keywords extracted by KeyBERT with the three most confident documents from each topic cluster into the LLM to generate more precise topic labels. Subsequently, documents with confidence scores below 0.8 or classified as outlier topics were treated as pending classification samples, using the optimized topic labels as classification labels. Finally, we utilized LLM fine-tuned with chain-of-thought reasoning to reclassify these samples, thereby enhancing the accuracy and consistency of topic division. The final optimized topic distributions and labels are presented in Table 3.

Table 3. Final Topic Modeling Results.

Topic	*Label*	*Count*
Topic 0	Citation diffusion	5
Topic 1	Citation measurement and prediction	8
Topic 2	Citation data management	5
Topic 3	Citation analysis across disciplines	4
Other	\	2

After completing the topic modeling, the study performed move recognition on the abstracts within each topic. Subsequently, the categorized move-specific texts for each topic were fed into the LLM for semantic integration and summarization, resulting in the generation of topic-based literature review modules. Finally, these modular reviews were synthesized into a comprehensive review following a "Topic–Move" structure. Figure 7 displays the input data and the LLM-generated literature review content for the "research problem" move under the "Citation Diffusion" topic. The results indicate that the generated literature review texts exhibit a clear focus on topics and strong coherence among documents within each topic. They effectively capture the research focal points and logical relationships in the literature, thereby

validating the effectiveness of the proposed methods in producing structured, high-quality literature reviews.

Template Filling Draft	Literature Review
1.1 Research issues related to "citation diffusion" Recent research issues on citation diffusion include the following: Study the role of citation range in developed economies and emerging powers on academic impact (especially in the social sciences), develop new citation range measurement methods, and explore the role of these ranges in improving academic impact. [1] Reveal the complex citation relationship and basic characteristics of retracted papers in molecular biology, including temporal and spatial patterns, reasons and publishers. [2] Analyze the similarities and differences in the knowledge base and diffusion trajectories of IPSUMGD and LDA to determine whether there are general patterns of successful innovation diffusion. [3] Study how policy information (policy design, topics and attributes) affects policy adoption. [4] Address the problem that existing methods are not suitable for measuring the preference for popular papers, and quantify this preference by calculating the popularity obtained in the references of papers. [5]	1.1 Research issues related to "citation diffusion" At present, citation diffusion research mainly focuses on identifying the knowledge base and its dissemination trajectory, quantifying the preference for popular papers, exploring the role of citation scope on academic impact and its measurement, the impact of policy information on policy adoption, and analyzing the citation relationship of retracted papers [1][2][3][4][5]. Through comprehensive discussion, these research issues are centered on measuring and understanding the role of citations in academic dissemination and influence, and provide a variety of methods for analyzing citations from diversified perspectives (such as innovation success, policy adoption, the impact of retracted papers, etc.) [1][3][5]. These studies are of great significance to the optimization of scientific research impact assessment methods, for example, exploring the academic impact of social sciences through new citation scope measurement methods [3].

Figure 7. Overview of Automatic Literature Review Generation Content.

Discussion

This study introduces an innovative method for automatic literature review generation that integrates both large and small models. It demonstrates unique advantages and significant potential in scenarios involving small-scale niche literature corpus. The proposed approach can be widely applied to periodic review tasks in niche fields, enhancing the efficiency of researchers in tracking the dynamic trends within their areas of study.

References

Agarwal, R., Singh, A., et al. 2024. Many-shot in-context learning. arXiv.

Asmussen, Claus Boye, and Charles Møller. 2019. "Smart Literature Review: A Practical Topic Modelling Approach to Exploratory Literature Review." *Journal of Big Data* 6 (1): 93.

Chen, Guo, and Tianxiang Xu. 2019. "Sentence Function Recognition Based on Active Learning." *Data Analysis and Knowledge Discovery* 3 (8): 53–61.

Han, Binglan, Teo Susnjak, and Anuradha Mathrani. 2024. "Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview." *Applied Sciences* 14 (19): 9103.

Vaishali, Ginni Sehgal, and Prashant Dixit. 2024. "A Comprehensive Study of Automatic Text Summarization Techniques." International *Conference on Emerging Innovations and Advanced Computing, Sonipat*, India 2024.

Wang, Yidong, Qi Guo, et al. 2024. "AutoSurvey: Large Language Models Can Automatically Write Surveys." arXiv.

Wei, Jason, Xuezhi Wan, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in L arge Language Models." *Advances in Neural Information Processing Systems* 35

Zhang, Tianyi, Varsha Kishore, et al. 2020. "BERTScore: Evaluating Text Generation with BERT." arXiv.

https://doi.org/10.51408/issi2025_135

Beyond Sentiment Analysis with ChatGPT: Classifying Authors' Perspectives on Russian Topics

Carolina Coimbra Vieira[1], Elena Chechik[2], Victoria Di Césare[3]

[1] *coimbravieira@demogr.mpg.de*
Max Planck Institute for Demographic Research, Max Planck Institute for Software Systems, and Saarland University (Germany)

[2] *elenachechik@gmail.com*
Europa-Universität Flensburg (Germany)

[3] *vdicesare@ugr.es*
University of Granada (Spain)

Abstract

In this study, we investigate whether fluctuations in the bilateral relations between the United States and Russia influence US authors' perspectives on Russian topics expressed in their research articles. Our analysis uses a dataset of approximately 14,000 Web of Science abstracts on Russia and Russian-related topics. We developed a methodology to annotate the abstracts as negative, neutral, or positive based on the author's perceived perspective on Russia and Russian topics. These categories are based on an ad hoc definition of positivity designed for this study, extending beyond conventional sentiment analysis. Based on this positivity definition, we use ChatGPT to annotate the abstracts and compare these annotations with results from traditional sentiment analysis methods. This approach provides a novel, annotated dataset that captures authors' nuanced perspectives on Russia and Russian topics.

Introduction

The relationship between the United States and Russia has long drawn international attention, marked by recurring fluctuations in diplomatic ties. In recent years, these tensions have escalated significantly, particularly in the context of the ongoing Russo-Ukrainian war (German, 2024; Oualaalou, 2021). Sentiment analyses of official US statements reveal a persistently negative stance toward Russia, increasingly framing the two nations as geopolitical adversaries (Berger Zalmanson, 2023). Beyond changes in official statements, shifts in diplomatic relations may also shape patterns of scientific collaboration (Li & Wang, 2024) and influence the tone in which countries are referenced in scholarly discourse.

In this paper, we focus on measuring the positivity in the perspective of authors towards Russia and Russian topics in their research articles over time. We address the following research question: *To what extent do fluctuations in US–Russia diplomatic relations influence the positivity expressed toward Russia in US-authored scientific abstracts?* We hypothesize that constructive and cooperative bilateral relations encourage US researchers to adopt a more positive perspective on Russian topics. Conversely, deteriorating relations may be associated with more negative portrayals of Russia in scientific discourse.

Our methodology consists of developing a nuanced approach to measuring authors' perspectives, beyond classical sentiment analysis. First, we conceptualize and operationalize the notion of "positivity" to capture the tone of authors' perspectives across a large corpus of scientific abstracts from the Web of Science (WoS) from 1990 to 2020. Second, we implement a mixed-methods approach that combines manual annotation with automated classification using ChatGPT to classify the positivity of the abstracts. Finally, we contribute a novel annotated dataset of US-authored abstracts that reflects how scholars have framed Russia and Russian-related topics across three decades. Additionally, we discuss technical limitations associated with ChatGPT-based annotation and propose directions for future research.

Data

Our dataset includes WoS articles about Russia and Russian-related topics from 1990 to 2020, compiled through a multi-step methodology detailed by Guba et al. (2024). We pre-processed the dataset to ensure that all the articles included the necessary information for our study, such as abstracts and affiliations. First, 38% of the articles did not have standard abstracts in the metadata structure and were removed from the dataset. Next, we categorized the articles in our dataset based on types of collaboration. For cases where an author's country of affiliation cannot be determined, we classify the collaboration based on the affiliations of the remaining co-authors. For 12% of the abstracts, we were unable to assign a country of affiliation for the co-authors and, therefore, were excluded from the analysis. Finally, we obtained our final dataset with 13.938 articles. Table 1 shows the number of articles in each category based on the co-authors' affiliation.

Table 1. Types of international collaboration based on the co-authors affiliations.

	Overall *(N=13,938)*
US alone	4,219 (30.3%)
Russia alone	1,963 (14.1%)
US + RU without other countries	264 (1.9%)
US + RU + other countries	56 (0.4%)
US + other countries, without RU	405 (2.9%)
RU + other countries, without US	466 (3.3%)
Other countries alone	6,565 (47.1%)

Methodology

To assess the perspectives of US researchers on Russian topics, we adopted an ad hoc categorization framework, classifying abstracts positivity as "negative," "neutral," or "positive" based on the authors' perspectives about Russia or Russian topics. A positive perspective emphasizes a favorable presentation of Russia or Russian topics under research with an optimistic tone. A neutral perspective emphasizes a balanced and impartial presentation of Russia or Russian topics under research, with a focus on stating facts and describing data. A negative perspective

emphasizes an unfavorable presentation of Russia or Russian topics under research with a critical tone. This framework was designed to capture nuanced perceptions rather than relying on traditional sentiment analysis. Figure 1 presents the annotation pipeline adopted in our methodology.

We randomly selected 1% of the abstracts (n = 140) for manual annotation by trained annotators. Three annotators annotated the abstracts independently using -1, 0, 1 to indicate whether the abstract positivity was "negative," "neutral," or "positive". The final annotation attributes the abstract as being related to a positivity when there was an agreement between at least two annotators for that positivity. The manually annotated subset served as a benchmark to validate ChatGPT's performance before using ChatGPT to classify the full dataset.

To annotate the subsample of 1% of the abstracts as well as the full dataset, we used the paid version of the ChatGPT API, employing the model "gpt-4o". We followed an ad hoc categorization framework for positivity and prompt, as detailed in the S1. The annotations produced by the ChatGPT model achieved an accuracy rate of 68% when benchmarked against the manual annotations.

Additionally, we compare the positivity annotations from ChatGPT with a traditional sentiment analysis approach. We classified all abstracts into three sentiment categories: negative, neutral, and positive. For this task, we used the paid version of the ChatGPT API (model "gpt-4o") to classify the abstracts according to the sentiment categories with a task-specific prompt, as presented in the S2.

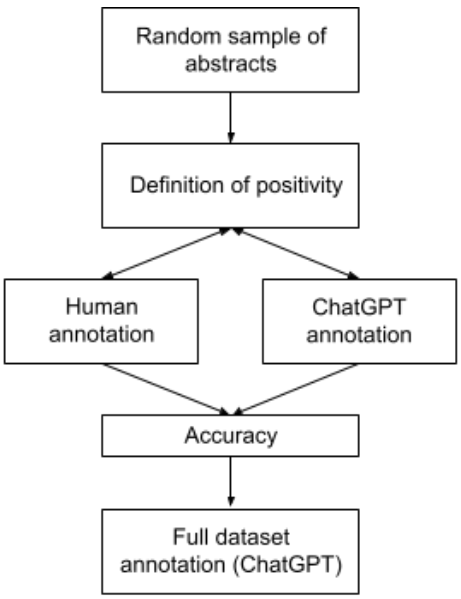


Figure 1. Abstracts positivity classification pipeline.

Preliminary results

Our preliminary results focus on describing the annotated dataset, which classifies authors' perspectives on Russia or Russian topics based on an ad hoc categorization of positivity. We also compare the positivity annotation with sentiment analysis.

Figure 2 shows the distribution of abstracts per year from 1990 to 2020 according to the positivity and sentiment classifications.

Figure 2a shows the absolute number of abstracts per year classified by ChatGPT API according to the author's positivity toward Russia and Russian topics. Overall, there is an increase in the number of abstracts over the years. Figure 2b shows the percentage of abstracts classified in each one of the positivity categories according to the ChatGPT API. While most of the abstracts are classified as neutral, 34% of the abstracts are classified as negative (see Table S3.1). The highest proportion of abstracts from the 90s and early 2000s are classified as negative.

According to the United States Department of State (2021), several key historical events defined US–Russia diplomatic relations from the 1990s to the early 2000s. In 1991, the US recognized the Russian Federation as the successor state to the Soviet Union. Another milestone was the establishment of the Bilateral Presidential Commission in 2009, aimed at fostering bilateral cooperation. Public opinion during this period also shifted. A Gallup survey (2025) reported that 60% of Americans held a favorable view of Russia in 1991, compared to 40% in 2009. However, when compared to our findings, a divergence between public sentiment and academic discourse becomes apparent. Despite high public favorability in 1991, our analysis shows a rise in the negativity of US research abstracts toward Russian topics. As shown in Figure 2b, between 1991 and 1994, approximately 40% of abstracts were classified as negative while in 2009 — when public favorability declined — the proportion of negative abstracts decreased to around 30%.

In contrast, 2014 marked a turning point in US–Russia relations, as the United States downgraded its political, economic, and military ties with Russia in response to Russia's violation of Ukraine's sovereignty and territorial integrity. This shift was followed in 2015 by a series of Western financial, administrative, and legislative sanctions that, alongside other factors, contributed to Russia's economic recession and the suspicion of cyber-interference activities in the 2016 US national elections (United States Department of State, 2021). Public opinion mirrored these tensions. By this period, favorable views of Russia among Americans had dropped to their lowest levels since the 1990s, with approximately 70% expressing an unfavorable opinion (Gallup Inc., 2025). In parallel, Figure 2b reveals a modest increase in the share of research abstracts classified as presenting a negative perspective on Russia-related topics after 2014. However, this increase did not reach the higher levels of negativity observed in the early 1990s and 2000s.

To evaluate the differences between our automated positivity classification and a standard sentiment analysis approach, we compared the results generated by ChatGPT based on our definition of positivity with those obtained through a conventional sentiment classification task. As shown in Figure 2c, the sentiment classification results from the ChatGPT API indicate that the majority of abstracts were labeled as neutral, with approximately 13% classified as negative. A detailed summary of the proportions of abstracts categorized as negative, neutral, and positive across all methods used in this study is provided in Table S3.1.

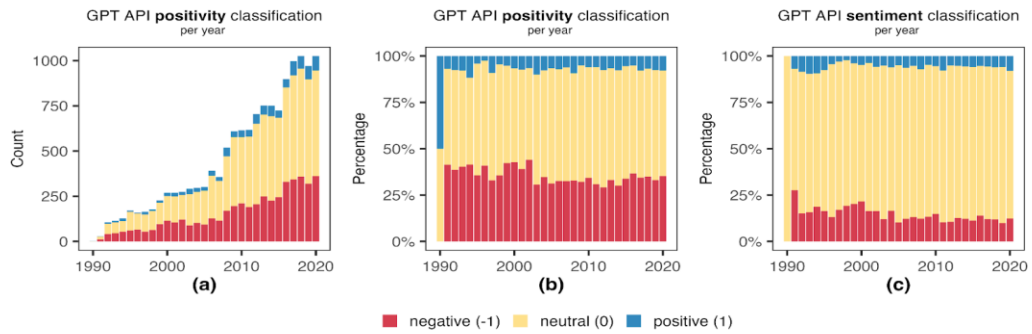


Figure 2. Yearly distribution of all the abstracts in the dataset according to the positivity and sentiment classification.

Additionally, Figures 3 and 4 illustrate the distribution of positivity scores in the abstracts disaggregated by area of study and co-authors' affiliation. As shown in Figure 3, abstracts co-authored by Russian scholars exhibit the highest proportion of negative positivity toward Russia and Russian topics. This trend is especially pronounced in papers exclusively authored by researchers affiliated with Russian institutions, which show a distinct peak in negative perspectives during the early 2000s. Figure 4 reveals notable disciplinary differences: fields such as Business, Economics & Management, Communication, History, Law, and Political Science display the highest prevalence of negative portrayals of Russia. However, temporal trends vary across disciplines. Political Science and Law consistently maintain high levels of negativity throughout the 1990–2020 period, whereas Business, Economics & Management, and History experience a decline in negative perspectives after the early 2000s. In contrast, Communication shows a marked increase in negativity, particularly in the years following 2010.

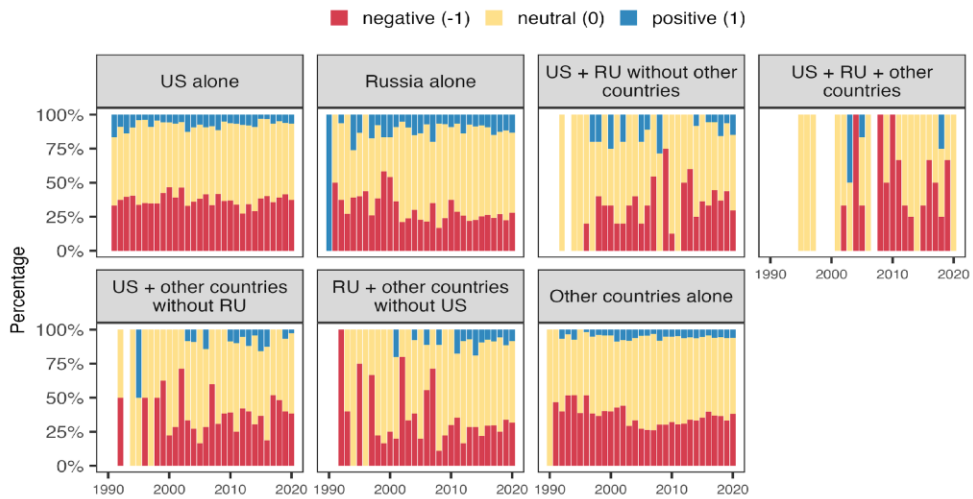


Figure 3. Yearly distribution of all the abstracts in the dataset according to the positivity by co-authors' affiliation.

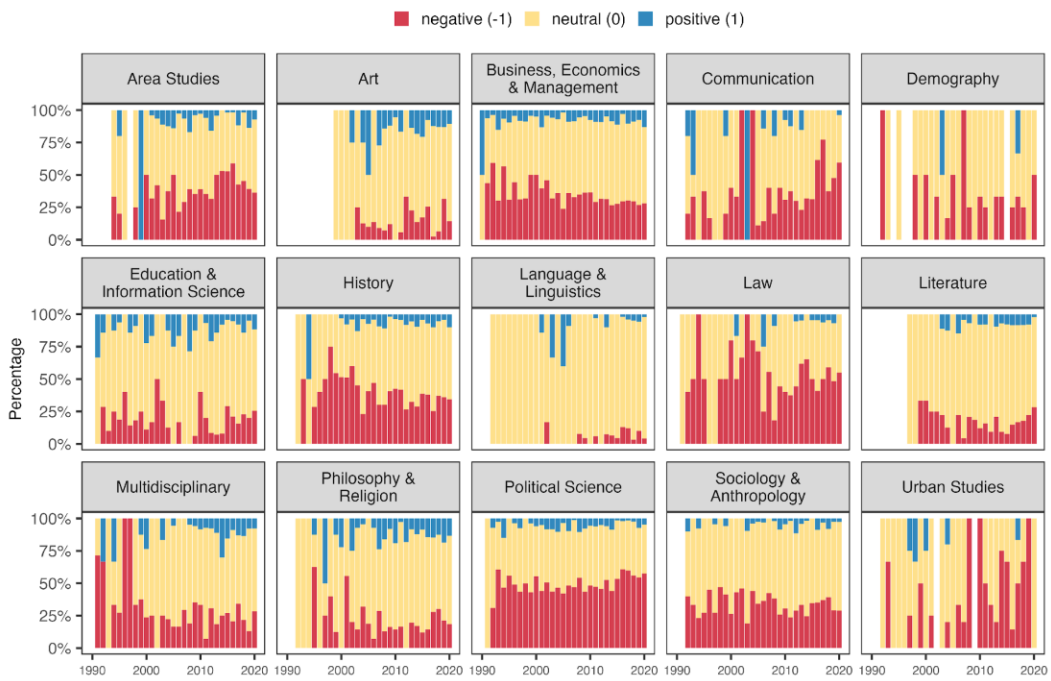


Figure 4. Yearly distribution of all the abstracts in the dataset according to the positivity by area of study.

Discussion

In this work, we proposed a methodological approach to address the technical challenge of classifying nuanced perspectives in research abstracts. Specifically, we focused on capturing the degree of positivity in US authors' perspectives on Russia and Russian-related topics by developing a methodology that goes beyond traditional sentiment analysis. By integrating manual annotation with automated methods using ChatGPT, we created an annotated dataset that not only facilitates the analysis of scholarly perspectives but also serves as a foundation for examining the influence of geopolitical relations on scientific discourse.

Our preliminary findings highlight key differences between our approach and conventional sentiment analysis techniques. While traditional sentiment classification categorized most abstracts as neutral, our ad hoc criteria enabled the identification of more subtle and context-specific perspectives toward Russia. These results demonstrate the potential of AI-assisted methods to capture more nuanced authorial viewpoints.

However, our methodology has limitations. While ChatGPT offers flexibility and general language understanding, it also exhibits inherent biases such as from its training on predominantly Western-centric data, as well as output variability and potential misalignment with the nuanced nature of geopolitical discourse. To address these limitations related to ChatGPT's on training and context, future work will compare ChatGPT's performance with models such as BERT, fine-tuned on policy texts and diplomatic corpora.

Regarding output variability, ChatGPT is a non-deterministic model, meaning it can produce different outputs when given the same input across multiple runs. In this study, we report the results from a single run of ChatGPT. Future work will include additional runs and confidence interval estimations to better understand the model's variability and reliability. This statistical approach will produce more robust measures of accuracy and uncertainty.

We also aim to contextualize shifts in authors' positivity toward Russia or Russian topics by incorporating key historical events into a year-by-year analysis. This will help identify geopolitical triggers that may influence scholarly perspectives. Finally, we will examine the composition of author teams by institutional affiliation to assess how collaboration patterns—such as heterogeneous versus homogeneous groups—are associated with the positivity expressed in the abstracts.

Supplementary materials

Supplementary materials are available at the link:
https://zenodo.org/records/15213176

References

Berger Zalmanson, D. (2023). *Can Sentiment Analysis Shed Light on International Relations? - A Case Study on United States Bilateral Interactions* [Master's Thesis]. University of Oxford.

Gallup Inc. (2025). *Russia*. Gallup.Com. https://news.gallup.com/poll/1642/Russia.aspx

German, T. (2024). From cooperation to confrontation: US-Russia relations since 9/11. *International Politics*, 61(3), 567–586. https://doi.org/10.1057/s41311-023-00524-x

Guba, K., Chechik, E., Tsivinskaya, A. O., & Buravoy, N. (2024). Global Ranking of Expertise about Russia. *Problems of Post-Communism*, 1–11. https://doi.org/10.1080/10758216.2024.2386995

Li, M., & Wang, Y. (2024). Influence of political tensions on scientific productivity, citation impact, and knowledge combinations. *Scientometrics*, 129(4), 2337–2370. https://doi.org/10.1007/s11192-024-04973-w

Oualaalou, D. (2021). *The Dynamics of Russia's Geopolitics: Remaking the Global Order.* Springer International Publishing. https://doi.org/10.1007/978-3-030-58255-5

United States Department of State. (2021). *U.S. Relations With Russia.* https://www.state.gov/u-s-relations-with-russia/

https://doi.org/10.51408/issi2025_147

Beyond Citations: Tracing and Validating the Rapid Adoption of AlphaFold in Biomedical Research Through Full-Text Analysis

Haochuan Cui[1], Yuzhuo Wang[2], Kai Li[3]

[1]*hcui94@hotmail.com*
School of Computing and Information, University of Pittsburgh, Pittsburgh, PA 15213 (USA)
Knowledge Lab, University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637 (USA)

[2]*wangyuzhuo@ahu.edu.cn*
School of Management, Anhui University, Hefei 230093 (P.R. China)

[3]*kli16@utk.edu*
School of Information Sciences, University of Tennessee, Knoxville, TN 37996 (USA)

Abstract

The emergence of AlphaFold, a deep learning model for protein structure prediction, has transformed biomedical research. This study analyzes full-text articles from the PubMed Central Open Access dataset to evaluate the dissemination and impact of AlphaFold. Focusing on 8,910 AlphaFold-related articles published between 2018 and 2023, we identify a significant rise in its application across major biomedical fields. Our analysis reveals discrepancies between mentions, citations, and actual usage: citation-based methods capture only 71% of articles mentioning AlphaFold in full text, while half of the articles citing foundational AlphaFold papers do not explicitly reference its name in the citation sentence. Despite being limited by the dataset's scope, this study highlights the need for advanced research methods and infrastructure to accurately assess the impact and usage of AI tools. Future work should explore a broader range of tools and datasets to provide a more comprehensive understanding of AI's influence on scientific research.

Introduction

The modern scientific landscape increasingly relies on advanced information technologies, including artificial intelligence (AI) and deep learning (Gao & Wang, 2024; Stevens et al., 2020). A prominent example of these technologies' transformative impact on science is AlphaFold. Developed by DeepMind in 2020, AlphaFold is a deep learning model designed to predict three-dimensional protein structures based on amino acid sequences (Ruff & Pappu, 2021). Initial testing demonstrated its exceptional accuracy (Jumper et al., 2021; Kovalevskiy et al., 2024), and it has since gained significant traction in fields such as data services, bioinformatics, structural biology, and drug discovery (Varadi & Velankar, 2023). The importance of AlphaFold was further underscored in 2024 when its developers received the Nobel Prize in Chemistry, marking a milestone in recognizing the profound influence of AI technologies on scientific research (Abriata, 2024).

In this project, we aim to comprehensively evaluate the impact of AlphaFold, as an emerging AI technology, on scientific research. AlphaFold provides an ideal case study due to its significant influence, as outlined above. However, assessing the impact of scientific software or algorithmic tools poses substantial challenges. First,

these tools may not always be cited or even mentioned in publications. Second, when cited, they may not consistently be represented by the same reference (Li et al., 2019). Consequently, relying solely on citation data to measure the impact of software and algorithms is widely recognized as inadequate (Howison & Bullard, 2016; Wang & Zhang, 2020). We argue that these methodological challenges have important implications for the growing research interest in AI for Science (Stevens et al., 2020). Addressing these issues requires the attention of researchers in scientometrics, research evaluation, and the science of science communities.

This short paper presents preliminary findings aimed at accurately evaluating the impact of AlphaFold. Rather than relying solely on citation data, we utilized full-text academic publications from the PubMed Central Open Access Subset (PMCOA) dataset. By examining the full text of academic publications and analyzing the contexts in which AlphaFold is mentioned, we aim to validate methodologies for tracing the impact of AI tools and develop a more nuanced understanding of how AlphaFold is utilized in scientific research. Specifically, this study seeks to address the following research questions:

RQ1: How has AlphaFold been disseminated in scientific research since its development?

RQ2: In what contexts is AlphaFold used in scientific research?

RQ3: How accurately do citations to AlphaFold papers reflect its impact and usage?

Our findings provide initial empirical evidence of AlphaFold's impact following its development and validate this impact by analyzing the contexts of name mentions. The results highlight the need to distinguish between citations, mentions, and usage of AlphaFold, as significant discrepancies exist among these measures. Furthermore, the findings challenge the validity of using (1) name mentions in titles and abstracts and (2) citations to key AlphaFold publications as proxies for its impact—a common practice in recent research (Hajkowicz et al., 2023; Liu et al., 2021). These insights call for a more nuanced approach to evaluating the influence of AlphaFold and other AI technologies in scientific research.

Methods

This study investigates the impact of AlphaFold on scientific research by analyzing the full-text content of academic papers, by taking the following major steps.

Data Collection

We downloaded a total of 609,615 full-text academic papers from the PubMed Central Open Access (PMCOA) dataset. Following the method proposed by Hsiao and Torvik (2023), we parsed the papers to extract key contextual information for each sentence, including section titles and citation details. Given the distinctiveness of "AlphaFold" as a term in scientific literature, we employed a dictionary-matching approach to identify sentences mentioning AlphaFold, including its key variations such as "AlphaFold" and "Alpha Fold." This process yielded a final dataset with 56,650 sentences from 8,910 papers published between 2018 and 2023 that reference

AlphaFold. To test the accuracy of this approach, we randomly selected 50 sentences from the dataset and 100% of them were the sentences that mentioned AlphaFold.

Entity Feature Identification

From the extracted sentences, we identified the following features for subsequent analysis:

1. **Section of the Sentence**: We analyzed the section of sentences as a signal for understanding how AlphaFold is utilized in scientific research. Previous studies have demonstrated that section titles provide valuable context for identifying the narrative function of sections, particularly within the IMRaD paper structure (Ma et al., 2022). Using a rule-based approach, we categorized section titles into six classical academic sections: Abstract, Introduction, Methods, Results, Discussion, and Others (Sollaci & Pereira, 2004). Keywords used to identify each section is available from our GitHub repository[1].
2. **Narrative Function of the Sentence**: We further leveraged a human-labelled dataset from Jurgens et al. (2018), which includes nearly 2,000 sentences annotated with one of five citation functions: *Uses*, *CompareOrContrast*, *Background*, *Extension*, *Motivation*, and *Future*. The definition of each category is also discussed in our GitHub repository. In this research, we are focused on the category of Uses, as it indicates that AlphaFold is used in the scientific research as a research tool. Using this dataset as training data, we fine-tuned the SciBERT model to classify the narrative function of sentences mentioning AlphaFold in our sample. We split the original dataset into three subsets: 1,600 samples for training, 200 samples for validation, and the remaining samples for testing. The fine-tuned model achieved an F1 score of 76%. To evaluate its performance on AlphaFold-related sentences, we applied the model and randomly selected 20 sentences for testing. Among these, six sentences were classified as 'Use,' and all were correctly identified. The remaining fourteen sentences were correctly classified as 'Not Use,' demonstrating the model's strong ability to distinguish 'Use' from other narrative functions.
3. **Research Areas of the Paper**: We identified the research topics of each paper in the dataset using Medical Subject Headings (MeSH) terms from the PubMed system. Each paper's topics were mapped to one of 16 top-level MeSH categories, representing broad research areas. For instance, the MeSH term "DiGeorge Syndrome," with tree number C16.131.077.019.500, belongs to category C (Diseases). Papers could be associated with multiple research areas.
4. **Representative References of AlphaFold**: We analyzed the references cited by the 8,910 papers mentioning AlphaFold in the PMCOA dataset. Using the PubMed Knowledge Graph (PKG) database, we retrieved all references cited in our final sample. In this preliminary research, we focused on the top three foundational references related to AlphaFold: Jumper et al. (2021), Mirdita et al. (2022), and Varadi et al. (2022).

[1] https://github.com/Wangyuzhuo95/ISSI2025

Results

Rapid Diffusion of AlphaFold in Academic Research

Figure 1(a) illustrates the rapid growth in the number of AlphaFold-related articles and citation sentences since 2018. Furthermore, we identified the top five research areas associated with AlphaFold using the MeSH system: (1) B (Organisms), (2) C (Diseases), (3) D (Chemicals and Drugs), (4) G (Phenomena and Processes), and (5) L (Information Science). Figure 1(b) shows the changing proportions of articles in each of these five areas over time. Notably, the shares of papers in categories D, B, and C have increased, indicating AlphaFold's growing use in applied research. In contrast, the share of papers in category L has declined, suggesting a relatively decreasing focus on analyzing and evaluating AlphaFold in technical literature, such as using AlphaFold to develop other tools and validating AlphaFold.

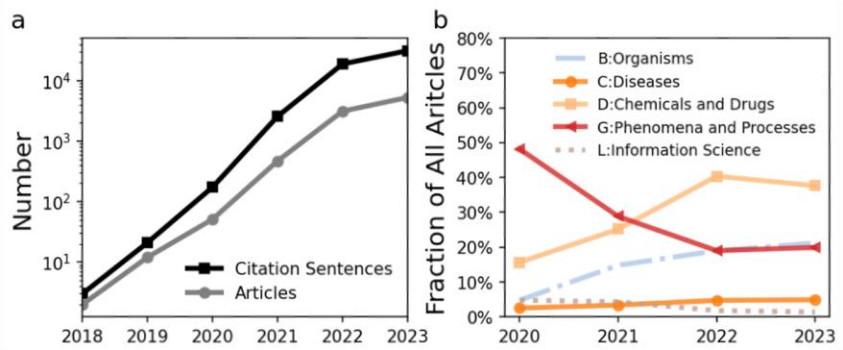


Figure 1. Rapid Diffusion of AlphaFold.

Inaccurate Impact by Traditional Methods

Using full-text academic publications, we compared citations and name mentions of AlphaFold to evaluate the accuracy of tracing its impact. We consider name mentions of AlphaFold as the gold standard for assessing its influence, given the known inaccuracies and limitations of software citations (Li et al., 2019).

Our analysis reveals that, of the 13,396 papers in the whole PMCOA dataset citing at least one of the three foundational references (many of these papers are not in our sample given that they did not mention AlphaFold in the text), only 51.0% explicitly mentioned AlphaFold in the text. Conversely, of the 8,910 papers mentioning AlphaFold, over 2,700 do not cite any of the three references, resulting in an accuracy of 71%. These findings indicate two key points: first, many papers cite key AlphaFold articles for purposes unrelated to AlphaFold, and second, relying solely on citations to trace AlphaFold's impact overlooks many relevant papers.

Figure 2 illustrates the proportion of articles citing the three foundational references (blue line) and those explicitly mentioning AlphaFold (including its variations) in the title or abstract. The proportion number are normalized by overall publication volume. These two measurements correspond to common approaches used to identify publications on the topic. Notably, no papers in our dataset cited the

foundational AlphaFold article (Jumper et al., 2021) before its publication in 2021. Additionally, we observe that the trends in both metrics remain consistent across the top five most prominent PMC domain fields, as shown in the supplementary figures.

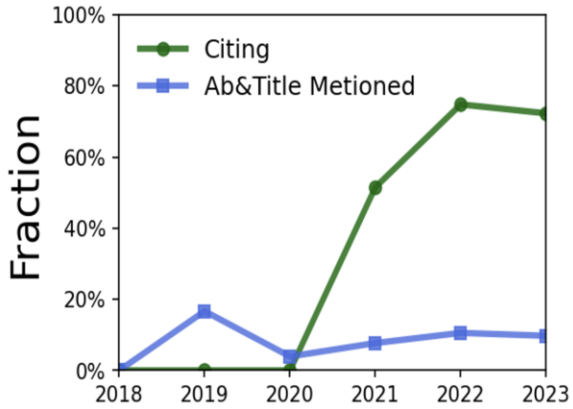


Figure 2. Share of all Articles Citing the Top Three Papers (green line) and Mentioning AlphaFold in the Title or Abstract (blue line).

Our findings carry significant methodological implications for empirical research on the impact of AI on science. Specifically, relying solely on citations or keyword searches in textual fields, such as titles and abstracts, is highly limited. These approaches often fail to capture all relevant articles, overlooking a substantial portion of related research.

For What Purposes is AlphaFold Mentioned in Papers?

To analyze the roles of AlphaFold in academic research, we used a BERT-based model to classify the narrative functions of each sentence into six categories. Figure 3(a) presents the raw counts of sentences for each category, broken down by narrative function and paper section, displaying only the top four categories for clarity. Figure 3(b) illustrates the percentage distribution of narrative functions within each section type. Our findings show that the predominant reason AlphaFold is mentioned in publications is its use in research. Nonetheless, other narrative functions are also frequently represented across the dataset.

a

Section	Uses	Compare& Contrast	Background	Others	All Types
Intro	926	366	2823	384	4499
Method	9933	335	644	151	11063
Result	7395	3599	1706	818	13518
Discussion	1327	1248	1937	1047	5559
All Sections	19581	5548	7110	2400	34639

b

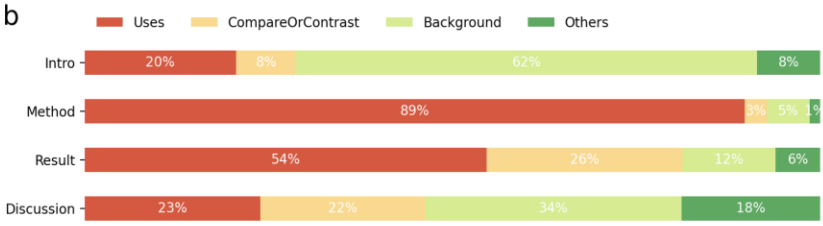


Figure 3. Distribution of Sentences Across Paper Sections and Narrative Functions.

Figure 4 illustrates the proportion of all articles containing at least one "Use" sentence related to AlphaFold. Our findings indicate that AlphaFold is increasingly utilized as a tool in the corpus. This observation aligns with prior evidence of an "instrumentalization" process for scientific tools within the citation landscape, which can be attributed to the need for such tools to undergo validation before being widely adopted (Li, 2021).

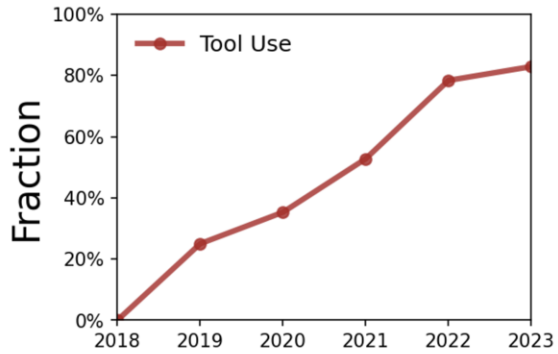


Figure 4. Fraction of Tool Use among AlphaFold-related articles (2018–2023).

Discussions and Conclusion

This paper presents preliminary findings from our project aimed at tracing the impact of AI technologies on science. Our analysis, focused on AlphaFold, highlights the rapid and transformative adoption of this deep learning model in biomedical research, as reflected in the PMCOA corpus. The adoption spans various research fields defined by MeSH terms, with a clear trend toward using AlphaFold in applied research rather than for other technical purposes (such as developing other tools and validating AlphaFold).

A critical insight from our study is the discrepancy between citations, mentions, and actual usage of AlphaFold. Traditional citation analyses often conflate these measures, leading to misunderstandings about the different types of impact associated with software and AI tools. Our findings show that citation-based methods capture 71% of articles mentioning AlphaFold in full text, and only half of the articles citing the three foundational AlphaFold papers explicitly mention AlphaFold within the paper.

These findings carry important implications for scientometrics, research evaluation and science of science research. As AI becomes an indispensable tool for a growing number of researchers, accurately evaluating its impact is an urgent priority for these communities. Our results underscore the significant limitations of relying on citation data and textual queries for assessing the impact of AI tools. These limitations highlight the necessity of full-text analysis for more accurate assessments. While recent studies have leveraged deep learning applications to identify AI technologies in publication texts (Gao & Wang, 2024), building robust data and methodological infrastructures to connect scientific publications to AI tools is essential for advancing this line of research.

In our next steps, we aim to systematically examine usage patterns of other biomedical technologies, such as CRISPR/Cas9. Comparing these patterns with those identified for AlphaFold will provide insights into whether similar trends are shared by other AI tools. This comparative approach will help us develop a more comprehensive understanding of the broader research landscape.

Acknowledgments

This work was partially supported by the National Social Science Fund of China (No.24CTQ027).

References

Abriata, L. A. (2024). The Nobel Prize in Chemistry: past, present, and future of AI in biology. *Communications Biology*, *7*(1), 1409.

Gao, J., & Wang, D. (2024). Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nature human behaviour*, 1-12.

Hajkowicz, S., Sanderson, C., Karimi, S., Bratanova, A., & Naughtin, C. (2023). Artificial intelligence adoption in the physical sciences, natural sciences, life sciences, social sciences and the arts and humanities: A bibliometric analysis of research publications from 1960-2021. *Technology in Society*, *74*, 102260.

Hsiao, T. K., & Torvik, V. I. (2023). OpCitance: Citation contexts identified from the PubMed Central open access articles. *Scientific Data*, 10(1), 243.

Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137-2155.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, *6*, 391-406.

Kovalevskiy, O., Mateos-Garcia, J., & Tunyasuvunakool, K. (2024). AlphaFold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, *121*(34), e2315002121.

Li, K. (2021). The reinstrumentalization of the Diagnostic and Statistical Manual of Mental Disorders (DSM) in psychological publications: A citation context analysis. *Quantitative Science Studies*, *2*(2), 678-697.

Li, K., Chen, P. Y., & Yan, E. (2019). Challenges of measuring software impact through citations: An examination of the lme4 R package. *Journal of Informetrics*, *13*(1), 449-461.

Liu, N., Shapira, P., & Yue, X. (2021). Tracking developments in artificial intelligence research: constructing and applying a new search strategy. *Scientometrics*, *126*(4), 3153-3192.

Ma, B., Zhang, C., Wang, Y., & Deng, S. (2022). Enhancing identification of structure function of academic articles using contextual information. *Scientometrics*, 127(2), 885-925.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature methods*, *19*(6), 679-682.

Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *Journal of molecular biology*, *433*(20), 167208.

Sollaci LB, Pereira MG. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J Med Libr Assoc*, 92(3), 364-7.

Stevens, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., & Brown, D. (2020). *AI for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science* (No. ANL-20/17). Argonne National Lab. (ANL), Argonne, IL (United States).

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439-D444.

Varadi, M., & Velankar, S. (2023). The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*, *23*(17), 2200128.

Wang, Y., & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of informetrics*, *14*(4), 101091

https://doi.org/10.51408/issi2025_134

Biblium: An Advanced Python Library for Bibliometric and Scientometric Analysis

Lan Umek[1], Dejan Ravšelj[2]

[1]*lan.umek@fu.uni-lj.si*, [2]*dejan.ravselj@fu.uni-lj.si*
University of Ljubljana, Faculty of Public Administration, Gosarjeva 5, SI-1000 Ljubljana (Slovenia)

Abstract

This paper introduces Biblium, a Python library designed to perform comprehensive bibliometric and scientometric analysis. Biblium replicates the core functionalities of the widely-used R package, Bibliometrix, while expanding its capabilities with several innovative features that enhance group analysis and visualization. This new tool aims to address the growing need for advanced, flexible, and reproducible bibliometric workflows within the Python ecosystem. A key distinguishing feature of Biblium is its ability to conduct group analysis, a functionality critical for analyzing bibliographic data that naturally divides into subsets based on factors such as publication periods, scientific disciplines, or geographic regions. Biblium implements multiple algorithms for comparative group analysis, enabling users to investigate associations between authors, keywords, sources, and other bibliometric elements across these subgroups. This group-level granularity facilitates insights into patterns of collaboration, thematic evolution, and differential impact across fields or timeframes. In addition to descriptive analysis, Biblium includes predictive algorithms that can forecast group membership based on key bibliometric indicators such as keywords, references, or citation patterns. This predictive modeling offers a forward-looking perspective on emerging research clusters and thematic areas. Biblium also empowers users to define custom concepts by leveraging keywords, abstracts, and other textual data, allowing for flexible and targeted analyses. The library features advanced visualizations, such as scatterplots that integrate multiple performance indicators (e.g., total number of citations, H-index, average year of publication, etc.) for authors, sources, and countries. This level of visualization extends the interpretative depth of bibliometric data and aids in more intuitive exploration and presentation of results. To support robust reporting and dissemination of findings, Biblium offers extensive options for exporting analysis outputs in formats including docx, html, tex, xlsx, and pptx. This versatility ensures seamless integration with academic publishing workflows and diverse dissemination platforms. Biblium also expands on traditional bibliometric indices by incorporating a variety of metrics beyond the H-index, providing a more nuanced evaluation of academic performance and influence. Biblium is expected to be publicly available on GitHub by June 2025, fostering an open-source community around bibliometric analysis in Python. This paper will detail the technical architecture, core algorithms, and key use cases of Biblium, demonstrating its application across different bibliometric scenarios. By providing a powerful, user-friendly alternative to existing tools, Biblium aims to accelerate bibliometric research across disciplines.

Introduction

In recent years, bibliometric and scientometric analyses have gained significant importance as essential tools for understanding trends, impact, and collaborations within scientific research. The increasing volume of scholarly publications and the need to assess the quality and influence of research outputs have made these analyses useful for researchers, policymakers, and institutions alike. For researchers, especially those who are new to a field, bibliometric tools provide a valuable overview of key trends, influential works, and potential collaborators, serving as a

crucial starting point for gaining insights and navigating the landscape of scientific literature.

As the demand for bibliometric insights grows, so does the need for robust tools that can process and analyze bibliographic data efficiently. Many software solutions and libraries have been developed to address this demand, with a notable example being the Bibliometrix library in R (Aria & Cuccurullo, 2017a). Bibliometrix provides a comprehensive suite of tools for bibliometric analysis, including descriptive statistics, network analysis, and visualization, making it a widely adopted resource for researchers across disciplines.

Parallel to the evolution of bibliometric tools, Python has emerged as a dominant programming language for data analysis and scientific computing. Its versatility, extensive library ecosystem, and ease of use have made Python the preferred choice for researchers and developers in various fields. Despite its widespread adoption, Python still lacks a comprehensive library that mirrors the functionalities offered by Bibliometrix. Moreover, existing software solutions often fall short in effectively analyzing (sub)groups of bibliographic data and leveraging data mining techniques for deeper insights. This gap underscores the need for innovative Python-based solutions to cater to the growing demand for bibliometric tools. This paper addresses that gap by introducing Biblium, a powerful Python library that brings advanced bibliometric analysis capabilities, replicates key functionalities of Bibliometrix, and introduces novel features—particularly for (sub)group analysis and data mining—to meet the evolving needs of the research community.

The paper is structured as follows: it begins with an overview of existing software solutions for bibliometric analysis, highlighting their capabilities and limitations, with a particular focus on Python-based libraries. This is followed by the introduction of Biblium. The final section discusses future directions and potential enhancements for Biblium, aiming to address current limitations and expand its utility for the bibliometric research community.

Software for bibliometric research

Bibliometric analysis relies on various software tools to process and visualize research data. This chapter provides an overview of key solutions, structured in two parts. The first section briefly introduces the most significant general-purpose bibliometric software, focusing on widely used and impactful tools rather than an exhaustive list. The second section delves into Python-based solutions, offering a detailed exploration of their functionalities, advantages, and implementation for bibliometric studies.

General-purpose bibliometric software

Various tools are available for bibliometric analyses, each with distinct strengths in processing, visualizing, and exploring data. VOSviewer (van Eck & Waltman, 2010) is widely used for its intuitive interface and visualization features. It creates bibliometric maps based on co-citation, co-authorship, and co-occurrence data. Compatible with databases like Web of Science, Scopus, and PubMed, it is particularly effective for visualizing large-scale networks and identifying research

clusters. The same authors implemented CitNetExplorer (van Eck & Waltman, 2017), a software solution which specializes in analyzing and visualizing citation networks. It allows interactive exploration of citation relationships and integrates with VOSviewer.

CiteSpace (Chen, 2006), focuses on trend analysis and detecting emerging topics. It identifies influential papers and key turning points in research fields using citation burst detection and network analysis, providing insights into the evolution of scientific domains.

Bibliometrix and its web-based interface, Biblioshiny (Aria & Cuccurullo, 2017b) offer comprehensive bibliometric capabilities. As an R package, Bibliometrix integrates science mapping, statistical analysis, and network exploration, while Biblioshiny provides an accessible interface for users without programming skills.

The Sci2 Tool (Team, 2009) supports advanced network and temporal analysis. It visualizes citation networks, collaboration patterns, and temporal trends. SciMat (Cobo et al., 2012) is tailored for longitudinal analysis and thematic evolution. It identifies research trends over time, focusing on knowledge progression and its influence across different periods. More detailed overview of best software solutions for bibliometric analysis can be found in (Moral-Muñoz et al., 2020).

Python libraries for bibliometric analysis

Metaknowledge (McLevey & McIlroy-Young, 2017) is one of the earliest Python-based tools for bibliometric analysis. The package offers various analytical capabilities, including longitudinal analysis, standard and multi-reference publication year spectroscopy, computational text analysis (e.g., topic modeling and burst analysis), and network analysis. One notable feature is its ability to estimate researcher gender by retrieving the Global Name Dataset from Open Gender Tracker's GitHub repository (*OpenGenderTracking*, 2013) and matching author and co-author names with probable genders.

Tethne (Peirson, 2016) is a Python-based tools for bibliometric analysis, developed to facilitate computational research in network science and bibliometrics. It was designed with a focus on co-citation, bibliographic coupling, and co-authorship analysis. Tethne provides functionalities for handling bibliometric data sourced from Web of Science (WoS) and Scopus. A key strength of Tethne is its integration with NetworkX, which allows users to analyze citation and collaboration networks effectively. However, its development has slowed down, and it lacks support for advanced natural language processing.

Pybliometrics (Rose & Kitchin, 2019) is a powerful Python library designed for bibliometric research with data sourced exclusively from Scopus. Unlike earlier tools, it offers direct access to Scopus API, allowing for large-scale data retrieval and analysis. Pybliometrics provides functions for citation counts, author productivity analysis, and institutional impact metrics. While it lacks built-in machine learning or NLP functionalities, it is used due to its efficient and programmatic approach to bibliometric research.

Scientopy (Ruiz-Rosero et al., 2019) is a relatively recent addition to the bibliometric analysis landscape. It offers comprehensive tools for analyzing bibliometric data

from WoS and Scopus, including citation analysis, co-authorship networks, and keyword trends. Scientopy is recognized for its ease of use and ability to generate detailed descriptive statistics and visualizations of scientific output over time. While it provides solid bibliometric functionalities, it does not incorporate sophisticated NLP techniques.

Litstudy (Heldens et al., 2022) was developed as an efficient Python package to assist researchers in conducting literature reviews. It supports the retrieval and processing of scientific metadata from multiple sources, including Scopus and other repositories. Its primary strengths lie in text mining and citation analysis, allowing users to extract key terms, identify trends, and map research landscapes.

TechMiner (Velasquez, 2023) is a Python-based graphical application tool that is useful for analyzing Scopus data by cleaning, renaming, and extracting relevant information while standardizing text formats. It offers analytical modules, including descriptive statistics, citation and co-word analysis, collaboration and conceptual mapping, term clustering, growth indicators, and impact assessments (H and M-index). Advanced tools like factor, correlation, latent semantic, and main path analysis enhance bibliometric insights. Additional features include thematic analysis, time-based tracking, top document ranking, and global visualization via a world map.

PyBibX (Pereira et al., 2025) is the most advanced Python library for bibliometric and scientometric analysis, incorporating cutting-edge artificial intelligence tools. It supports data from Scopus, WoS, and PubMed, providing comprehensive exploratory data analysis (EDA), citation, collaboration, and similarity networks. A major innovation of PyBibX is its AI-driven capabilities, including embedding vectors, topic modeling, and text summarization. It integrates models such as Sentence-BERT, BERTopic, BERT, chatGPT, and PEGASUS to enhance bibliometric insights. PyBibX stands out as the first bibliometric tool to feature AI-driven conversational analytics, allowing researchers to interact with bibliometric results dynamically.

None of these Python libraries can analyze groups of documents from bibliographic datasets, a key limitation that differentiates them from more advanced bibliometric tools. While some, like Metaknowledge, Tethne, and Pybliometrics, provide functionalities for network analysis and citation metrics, their capabilities are restricted to specific tasks such as co-citation analysis or institutional impact assessment. Many of these tools, except for TechMiner and PyBibX, lack advanced NLP (and AI-driven analytics), making them significantly less comprehensive than Bibliometrix in R.

Bibliometric analysis with Biblium

Biblium is a tool designed for the analysis and visualization of bibliographic data. In this section, we will describe its basic functionalities, showing how it can assist in exploring and interpreting scientific publications. To demonstrate its capabilities, we will use a sample dataset comprising the 500 most-cited documents related to bibliometric, scientometric, and informetric research. This dataset serves as an illustrative example to highlight Biblium's functionalities.

Currently, Biblium is hosted on a private GitHub repository, accessible to selected collaborators for review and testing. Plans are in place to make the repository publicly available by June 2025 on the site: https://github.com/lan-umek/biblium

Python Libraries for Biblium

In Biblium, a variety of Python libraries were employed to handle bibliographic data, perform statistical analyses, and generate visualizations. The primary data structure used was the Pandas DataFrame, which served as the main object for storing bibliographic data and output dataframes, such as performance measurements and scientific production metrics (Mckinney, 2010). Pandas was extensively utilized for data transformations, including sorting, filtering, and computing new features, while NumPy provided foundational support for numerical operations (Harris et al., 2020). For statistical analysis (statistical tests, clustering, entropy calculation), Scipy.stats was employed (Virtanen et al., 2020), and predictive modeling tasks, such as logistic regression, were conducted using Scikit-learn (Pedregosa Fabian et al., 2011) and Statsmodels (Seabold & Perktold, 2010).

Visualization is important aspect of Biblium, with Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) being used for creating plots, Plotly for k-field plots and geographical maps (Inc., 2015), Squarify for treemaps (Laserson, 2009), and upsetplot (Nothman, 2023).

Network analysis and visualization were achieved using NetworkX (Hagberg et al., 2008), complemented by CDlib (Community Detection Library) for partitioning algorithms (Rossetti et al., 2019). Additionally, Pyvenn (Tctianchi, 2014) was used for generating Venn diagrams, while adjustText (Flyamer, 2012) ensured clear labeling on graphs by preventing overlap and properly positioning labels.

Text processing tasks, such as lemmatization, stop words removal, and n-gram analysis, were handled by Natural Language Toolkit NLTK (Bird et al., 2009), while Gensim (Rehurek & Sojka, 2011) was employed for topic modelling. For image manipulation, the PIL library was utilized (Umesh, 2012), and Wordcloud (Mueller, 2010) was used to generate word clouds.

Initialization

The provided code snippet demonstrates the initialization process of Biblium's core BibliumAnalysis class.

```
>>> import biblium as bb
>>> ba = bb.BiblioAnalysis(f_name="data.csv", db="scopus")
```

The BiblioAnalysis class in Biblium can be initialized with either a file name (e.g., "data.csv") or a Pandas DataFrame. The db argument, specifying the database source, is currently limited to "wos" (Web of Science) and "scopus", with support for additional databases expected to be added soon. Biblium supports input files in csv, xlsx, and txt formats, and plans are in place to accommodate more file types in the future.

Although db is a keyword argument and not strictly required during initialization, failing to provide this information will result in a BiblioAnalysis instance with almost no meaningful or interesting results computed. Providing the correct database information is essential to unlock the full potential of the tool. Biblium adopts the terminology used in the Scopus database. If data from a different database are provided, Biblium automatically adjusts column names to align with the closest matching Scopus terms.

In the initialization process user can change several additional keyword arguments, the most important being the pre_compute argument and res_folder. The pre_compute parameter controls the level of preprocessing and descriptive statistics performed during initialization. Setting pre_compute=1 computes basic descriptive statistics, while increasing the value up to pre_compute=4 adds progressively more advanced computations, including lemmatization of abstracts, the computation of new derived features, and other detailed analyses.

When the res_folder parameter is provided, a directory is created with structured subfolders: networks, plots, tables, and reports, where results are saved in various formats. This ensures an organized structure of the results.

In addition to these core parameters, various other optional parameters can be customized during initialization. These include settings for the default color scheme, the resolution of png plots to be saved, the language of the output, and more. While these parameters are not described in the paper, they will be fully detailed in the Biblium tutorial and documentation available on GitHub. This ensures users have the flexibility to adapt Biblium to their specific needs and preferences.

Main Information

By calling

```
>>> ba.get_main_info()
```

two dataframes are computed: main_info_df and production_df. The main_info_df includes basic publication metrics such as the total number of documents, sources, and citations, as well as details like the number of documents per source and the proportion of cited documents. Temporal aspects include ranging from the first and last years of publication, the period length, and most productive year, to other statistics like the average year, standard deviation of publication year, and quartiles of publication years (Q1, median, Q3). Growth trends are quantified through metrics such as overall document growth and growth rates over the past year, five years, and ten years. Citation analysis includes the highest number of citations for a document, the H-index, and G-index, alongside average citation metrics for all and cited documents specifically. The analysis also highlights productivity by most frequent sources, countries, and keywords.

Authorship and collaboration metrics enrich the analysis with insights into the number of unique authors, documents per author, and co-authorship patterns, including single-author and multi-author documents. These are further detailed through collaboration indices and international collaboration trends. Reference data

provides an understanding of citation behavior, including references per document, correlation between publication year and average reference year, and the distribution of unique references. Language and accessibility metrics describe the most frequent document languages, the number of multi-language publications, and open-access availability.
The production_df data frame captures both annual and cumulative statistics related to scientific output (number of documents, total number of citations). It details the number of documents published each year, allowing researchers to observe patterns of growth or decline in productivity over time.
The scientific production is plotted using the

```
>>> ba.plot_production(cut_year=2000)
```

which includes a cutoff at the year 2000 to reduce the wide spread of data and create a more visually representative graph. The plot shows the yearly number of documents as bars, alongside a line representing the cumulative number of citations over time (Figure 1).

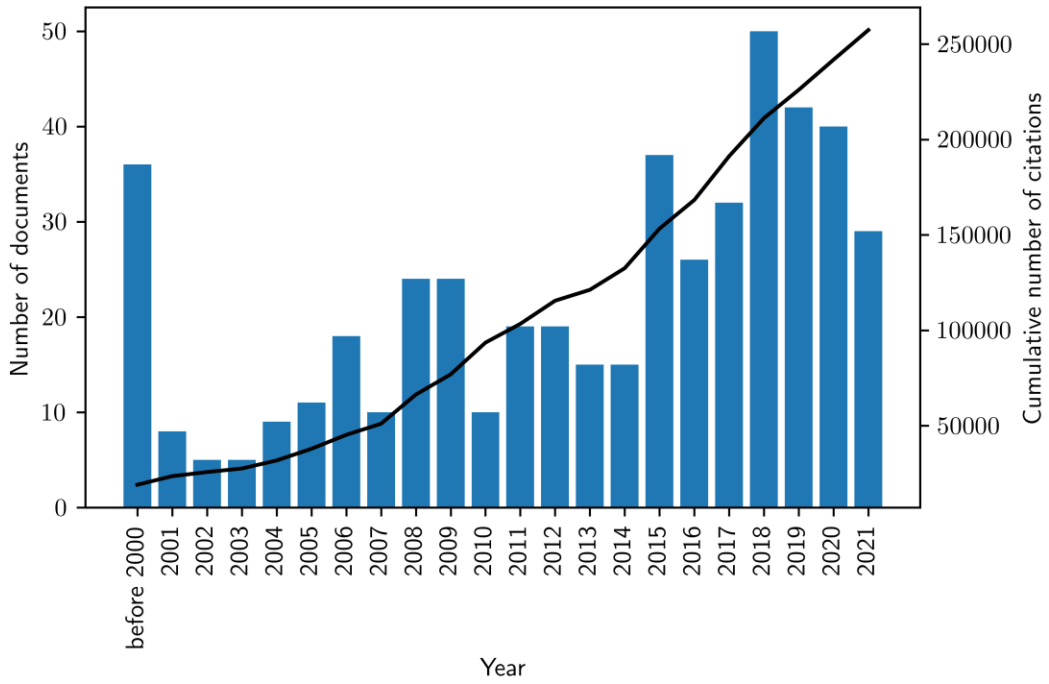


Figure 1. Scientific production plot from Biblium: annual number of documents and cumulative citations over time, with a cut-off year applied for improved visualization.

Measuring Performance

In Biblium, performance indicators are implemented to analyze scientific production across various units of observation, such as sources, authors, countries of corresponding authors, references, keywords, scientific fields, etc. These functionalities allow users to explore patterns and trends in scholarly output by focusing on specific dimensions, like the distribution of contributions by country, the prominence of particular keywords, or the impact of sources and references.

The process involves two stages. First, Biblium counts the number of occurrences for each unit of observation. Next, these counts are used to compute performance indicators for a user defined subset of units. The following snippet is implemented to count the occurrences of sources and compute additional statistics for the top 20 sources based on the number of documents (Users can define subsets based on specific criteria, utilize subsets from other dataframes, or even employ regular expressions for the selection.). This function takes additional parameter "level", ranging from 0 to 4, which determines the extent of the computed statistics. The selection of items for statistical analysis in Biblium is impressively versatile. At level 0, only the source counts are returned. Higher levels progressively compute more detailed metrics, level 4 includes all implemented statistics, offering a comprehensive analysis of source performance. This flexible approach allows users to tailor the depth of analysis to their specific needs.

```
>>> ba.count_sources()
>>> ba.get_sources_stats(top=20, level=1)
```

At level 1, the code computes essential statistics, including total citations, the average publication year, and the H-index, providing a solid foundation for evaluating source performance. Level 2 expands the analysis with additional metrics, such as the G-index, cumulative citation counts (C5, C10, ..., C100), and publication year distribution statistics like the first quartile (Q1), median, and third quartile (Q3). At level 3, the code computes the interdisciplinarity by calculating the normalized entropy of counts related to different scientific disciplines, offering insights into the diversity of a source's contributions. Level 4 focuses on advanced or specialized indices, such as the HG-index (Alonso et al., 2010), m quotient (Hirsch, 2005), Tapered H-index (Anderson et al., 2008), A-index and R-index (Jin et al., 2007), and q^2-index (Cabrerizo et al., 2010). These metrics cater to nuanced evaluation needs and are ideal for more sophisticated analyses. Many other advanced indices will be introduced in future updates, enhancing the scope of Level 4 statistics and providing users with more comprehensive analytical tools. The code is designed to be general and adaptable, extending beyond the evaluation of sources to other units of observation, such as authors, countries of corresponding authors, references, keywords, and scientific fields.

The computed statistics are stored in a pandas DataFrames named sources_counts_df and sources_stats_df, which provides a structured format for further analysis and visualization. This DataFrame can be easily leveraged for various plotting purposes, enabling users to explore the data visually. One of the most insightful visualizations

is a scatterplot, as it can incorporate up to four indicators simultaneously. In Biblium it can be plotted using the default parameters by calling

```
>>> ba.scatter_plot_top_sources()
```

By default, the scatter_plot_top_sources function visualizes the number of documents on the x-axis and total citations on the y-axis, both on logarithmic scales. The H-index determines marker size, while the average publication year defines marker color, and abbreviated source titles serve as labels. Additionally, one categorical property can be shown with different shapes of the dots. The plot includes the top 20 sources based on the number of documents. Optional features like mean lines (dashed line indicating the means of variables on both axes), mean values, and regression lines are disabled by default. Users can further customize annotations with arrow properties, ensuring the visualization is both insightful and adaptable to different datasets. The scatterplot is saved in the "plots" folder in three different formats: png, pdf, and svg. For illustration, a scatterplot produced on the dataset used in the paper is shown in Figure 2.

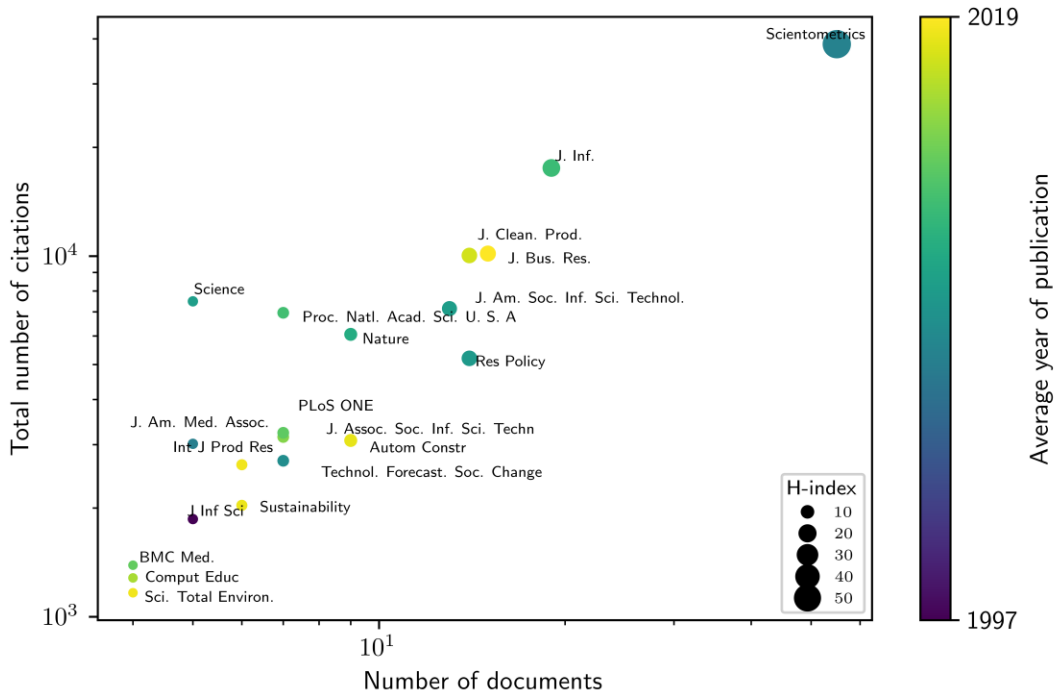


Figure 2. Scatter plot of top 20 sources from Biblium: total number of documents versus total citations on logarithmic scales, with marker size indicating H-index and color representing the average publication year.

Networks

Biblium includes several types of bibliographic networks to facilitate the analysis of scientific literature. These networks encompass keyword co-occurrence, co-authorship, co-citation, co-occurrence of custom-made concepts, and bibliographic coupling. However, a citation network between documents is missing. For illustration, a keyword co-occurrence network will be presented, based on authors' keywords, to demonstrate thematic structures within the dataset. The keyword co-occurrence network can be computed using the snippet

```
>>> ba.get_keyword_co_net()
```

First, the network is computed and saved as a .net file in Pajek (Batagelj & Mrvar, 2004) format (in network subfoder of the results folder). Then, partitioning (community detection, by default using the Louvain algorithm (Blondel et al., 2008)) is applied to the nodes. The partition is saved in a .clu file (Pajek format). Additionally, several vectors (numerical properties of the nodes) are computed, including the number of documents, total number of citations, average year of publication, and H-index. These are saved in .vec files (Pajek format). The network can then be plotted in Pajek or any other software that supports the Pajek format. Biblium can also plot it. By calling

```
>>> ba.plot_keyword_co_net()
```

the network and overlay representations are plotted: in the network view, color represents the partition, and size corresponds to the number of documents; in the overlay view, color indicates the average year of publication, and size represents the number of documents. Notice that the user can specify the number of keywords to include, the partitioning algorithm, and the network layout (default: spring_layout from NetworkX). The plots are saved in png, pdf and svg format.

Besides the classical keyword co-occurrence network, other visualization methods are available. A heatmap can be generated, which provides a matrix representation of keyword co-occurrence frequencies, highlighting patterns and relationships in a structured way. Additionally, a thematic map can be generated to represent clusters of keywords within a two-dimensional space. By default, the x-axis corresponds to centrality, while the y-axis represents density. This visualization facilitates the identification of thematic structures, including motor themes, highly developed themes, as well as emerging or declining themes and basic themes, based on their positioning within the quadrants. The Figure 3 represents a “classical” keyword co-occurrence network of the top 20 keywords.

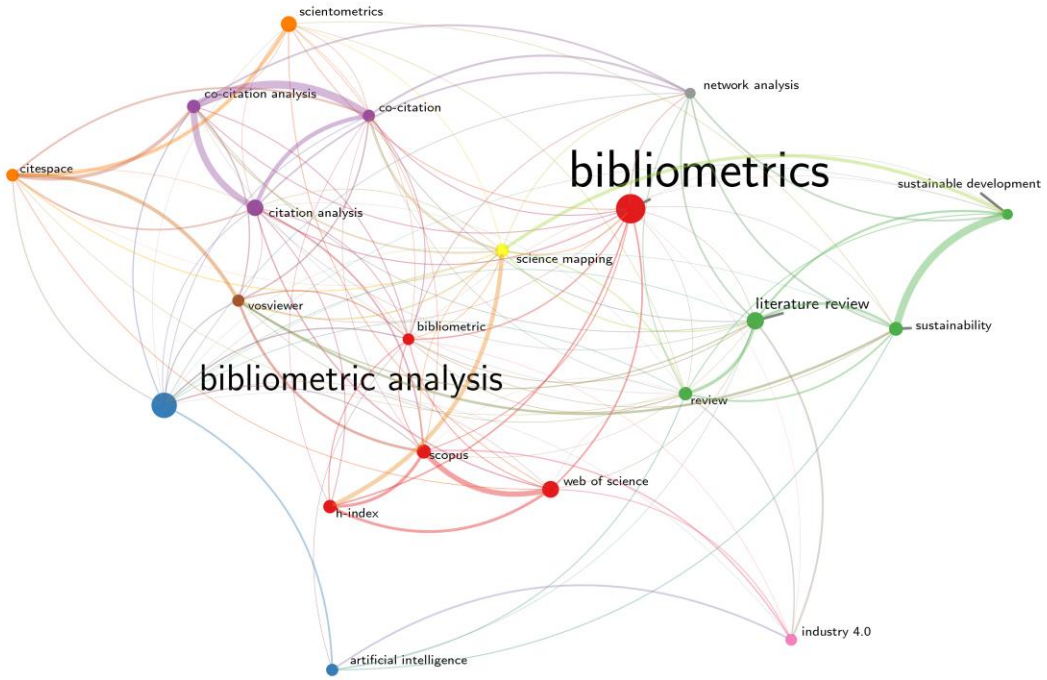


Figure 3. Co-occurrence network of author's keyword visualization from Biblium: direct output displaying relationships between key terms, with no synonym cleaning applied.

In the Figure 3, some keywords share the same meaning, which can create redundancy. For example, "bibliometric analysis" and "bibliometrics" refer to the same concept, just as "citation analysis" can be simplified to "citation," and "systematic review" to "review." To address this, Biblium supports keyword cleaning by merging synonyms and removing irrelevant terms.

Users can provide a dictionary mapping of synonyms to a preferred term and a list of words to be removed. Alternatively, they can prepare two Excel files (.xlsx) containing this information in a structured format. The structure of the synonym file can follow one of these formats: Column-based: The column name represents the term that should be kept, while all words listed below it will be replaced by this term; or Row-based: The first word in each row is the preferred term, and all other words in the same row will be replaced by it. Biblium also offers automatic cleaning by standardizing keywords, such as converting plural forms to singular where appropriate. This feature ensures a cleaner, more consistent keyword representation in the network. The example of keyword co-occurrence network after the cleaning is shown in Figure 4.

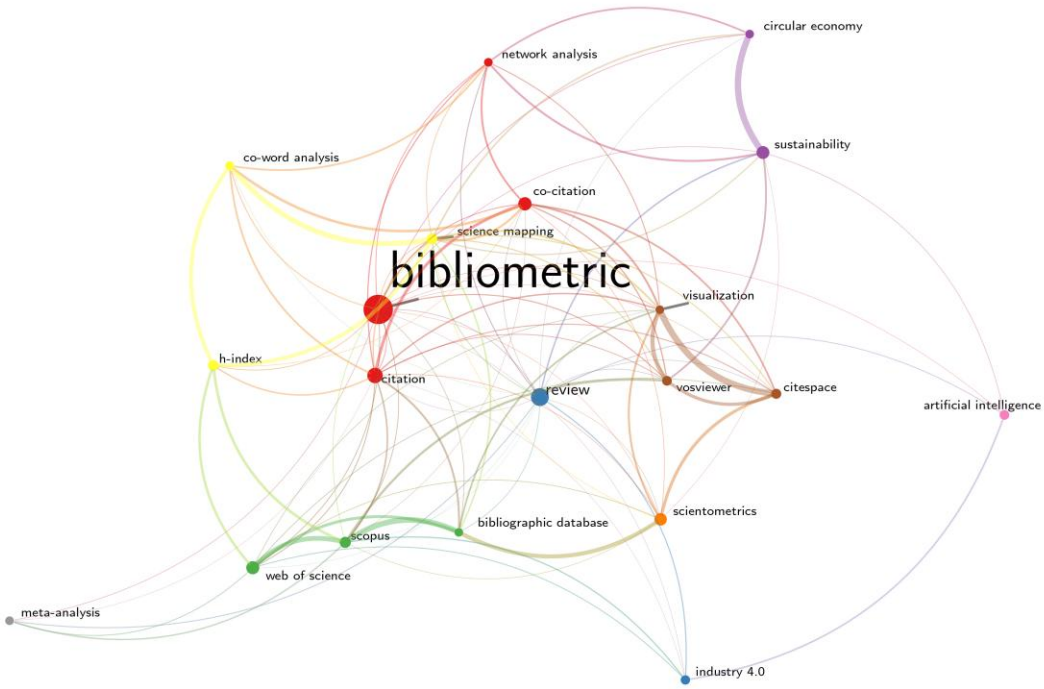


Figure 4. Co-occurrence network of author's keyword visualization from Biblium: cleaned output with synonyms consolidated.

Reports

Biblium supports comprehensive export capabilities that allow users to generate professional bibliometric reports in multiple formats. Reports can be exported as Excel (.xlsx) files with styled tables, as well as Word (.docx) and PowerPoint (.pptx) documents, following a user-defined structure stored in Excel templates. For Word exports, Biblium uses the python-docx library, enabling detailed styling, structured headings, captions, and dynamic table and figure generation. PowerPoint reports are built using python-pptx, allowing for custom slide layouts, embedded figures, and narrative content aligned with analytical outputs. These flexible export options ensure seamless integration with dissemination workflows and enhance the clarity of bibliometric insights. Reports can be generated by calling

```
>>> ba.save_reports(formats=["docx", "xlsx", "pptx", "tex"], f_name="report")
```

where the user specifies in which format(s) the report should be saved and under what file name.

Other functionalities

In addition to the core functionalities discussed in detail throughout this section, Biblium offers several other useful features that enhance bibliometric analysis. These functionalities provide additional flexibility and depth for users interested in refining

their research. For instance, the tool supports the computation of new variables by integrating information from abstracts, titles, and keywords. It also enables users to define custom concepts using keywords and regular expressions, facilitating the identification of more complex thematic structures. Preprocessing tools such as lemmatization, stopword removal, and user-defined word filtering further improve text clarity for subsequent analysis.

Beyond text processing, Biblium includes a range of scientific metrics that capture interdisciplinary connections, reference statistics (sources, authors, age distributions), and top cited documents and references. Users can segment data into custom time periods for group analyses, track the dynamics of sources, authors, and keywords, and explore trending topics over time. Additionally, various statistical techniques allow for association analysis between general concepts (such as keywords, authors, and sources) and user-defined categories, offering deeper insights into the analysed dataset.

Biblium includes several clustering algorithms for grouping the documents and other units, such as sources and authors, etc. These algorithms rely on keywords, references, or any user-defined (dis)similarity measure. The main clustering methods implemented are hierarchical clustering, k-means, and bibliographic coupling. Once cluster membership is determined, statistical evaluation—such as comparing groups using descriptive statistics and statistical tests—can be performed. Additionally, data mining approaches like logistic regression can be applied for further analysis. However, group membership does not necessarily have to result from a clustering approach, as statistical comparison and data mining techniques can also be used independently to analyze predefined groups.

For more advanced exploration, Biblium includes topic modeling (Latent Dirichlet allocation, LDA (Blei et al., 2003)), factor analysis (correspondence analysis, hierarchical clustering, multidimensional scaling, MDS), sentiment analysis, and extended visualizations. The k-field plot, an extension of the traditional three-field Sankey diagram, provides an effective way to relate multiple concepts within a dataset. Visualization tools such as bar plots, lollipop plots, violin plots, heatmaps, Venn diagrams, word clouds, and treemaps make it easier to interpret results. Additionally, classic bibliometric models, including Lotka's law (author productivity) and Bradford's law (source dispersion), are implemented to provide theoretical context.

Biblium ensures the preservation and reusability of computed data through pickling a BiblioAnalysis object and the storage of indicator dataframes, making it easy to retrieve results for further analysis. While these functionalities are not explored in full detail in this section, they provide additional opportunities for users to deepen their bibliometric investigations and tailor analyses to their specific research needs.

Bibliometric group analysis with Biblium

Group analysis plays an important role in bibliometric studies by enabling the identification of patterns and trends within specific segments of a dataset. Groups can be defined based on predefined criteria such as time periods, geographic regions, or scientific disciplines, but they can also be formed using clustering algorithms,

which categorize documents based on keywords, references, or abstracts (e.g., bibliographic coupling). The group analysis offers a different and more detailed bibliometric analysis than analyzing the whole dataset. For example, a temporal group analysis can reveal how scientific priorities have changed over decades, while a geographic analysis can highlight the research strengths of different countries or institutions. Similarly, grouping by scientific disciplines allows us to examine how different fields contribute to the overall scientific landscape and interact with one another, which topic are more associated with particular domains, etc.

In this chapter, we will illustrate the functionalities of Biblium on group analysis based on scientific disciplines, using classification data extracted from the Scopus webpage. Each document is assigned to one or more categories—Social Sciences, Physical Sciences, Health Sciences, or Life Sciences (excluding Multidisciplinary for this example)—allowing us to explore discipline-specific trends and contributions. We will demonstrate how to initialize and conduct this analysis within Biblium.

In Biblium, group-based analysis is performed by initializing an object of the class BiblioGroup. This initialization builds on the standard setup used in BiblioAnalysis, with the key addition of a flexible parameter called group_desc, which defines how documents are grouped for analysis. The group_desc can take multiple forms, offering extensive flexibility: (1) the name of a column in the bibliometric DataFrame, where each value indicates the group membership of a document; (2) a binary indicator DataFrame, with rows representing documents and columns representing groups; (3) the name of a multi-valued column (e.g., keywords or authors), where items are separated by a delimiter; or (4) a dictionary of regular expressions, used to classify documents based on text in a specified column (e.g., abstracts). When time-based grouping is required, group_desc="Year" can be combined with cutpoints or n_periods to define custom or evenly spaced time periods. Biblium automatically detects the appropriate grouping logic and constructs a binary group matrix that supports both disjoint and overlapping group structures. This design ensures accurate and scalable bibliometric analysis across a wide range of use cases.

```
>>> bg = bb.BiblioGroupAnalysis(f_name="data.csv", db="scopus", group_desc=”Sciences”)
```

Notice that the initialization process is similar to that of the BiblioAnalysis object, with the key difference being the inclusion of the group_desc parameter. This parameter corresponds to a column in the dataset (data.csv) that contains information about the scientific classification of each document. If a document is associated with multiple scientific fields, they are separated by a semicolon (;). This classification column is not included in the dataset directly downloaded from Scopus and must be generated by the user. However, for datasets originating from Scopus, Biblium provides specialized functions to assist in adding this classification column. These functions require additional metadata about sources, which is accessible on the Scopus webpage for registered users.

Main Information

BiblioGroupAnalysis extends the functionalities of bibliometric analysis to grouped data, ensuring that most features available for individual datasets are also implemented for groups. Key analyses such as main information, scientific production, and performance indicators are all accessible within this framework, maintaining a consistent output structure. Typically, results are stored in pandas DataFrames, which can be conveniently saved in xlsx format for further use. However, in BiblioGroupAnalysis, outputs are separated by group, allowing for comparative insights across different categories. The following snippet

```
>>> bg.get_main_info()
```

computes the main information for grouped data, and the selected rows of the output are displayed in the table 1.

Table 1. Biblium summary statistics (part of the computed dataframe): overview of document distribution, sources, citations, and collaboration trends across the groups (scientific disciplines).

	Social Sciences	*Physical Sciences*	*Health Sciences*	*Life Sciences*
Number of documents	295	265	80	39
Number of sources	121	107	65	35
Timespan	1983: 2022	1976: 2022	1996: 2022	1997: 2022
Total citations	164,203	143,355	35,392	17,933
H-index	248	239	80	39
Average year of publication	2012.61	2012.95	2010.91	2014.59
Top 5 sources	Scientometrics, Journal of Informetrics, Journal of Business Research, Journal of Cleaner Production, Research Policy	Scientometrics, Journal of Informetrics, Journal of Cleaner Production, Journal of the American Society for Information Science and Technology, Automation in Construction	JAMA, BMC Medicine, BMJ (Online), Environmental Research Letters, PLoS Medicine	PLoS Biology, Expert Opinion on Biological Therapy, PeerJ, Trends in Ecology and Evolution, Agronomy
Top 5 keywords	bibliometrics, bibliometric analysis, literature review, web of science, citation analysis	bibliometrics, bibliometric analysis, web of science, literature review, citation analysis	bibliometrics, bibliometric analysis, open access, scientific publishing, citation analysis	bibliometric analysis, scientometrics, citation analysis, bibliometrics, regenerative medicine
Collaboration index	3.51	4.21	5.42	5.12

These kinds of tables allow for comparisons across the analyzed groups in terms of research output, impact, collaboration as well as the content. Differences in the

number of documents and sources highlight variations in publication density and dispersion, while the timespan and average publication year indicate dynamics of the research trends. Citation metrics, including total citations and H-index, reveal differences in research influence and scholarly impact. The top keywords and source illustrate publishing trends and possible overlap between groups. Note that the presented statistics represent only a subset of the possible results and are provided here for illustrative purposes. To visually represent the top sources for each group (in this case, scientific disciplines), we can use the following snippet:

```
>>> bg.plot_top_sources_barh()
```

The output is shown in Figure 5. The plot shows the distribution of documents across sources, categorized into four groups based on scientific discipline: Social Sciences, Physical Sciences, Health Sciences, and Life Sciences. It is evident that some sources contribute to multiple scientific disciplines. Similar graphs can be generated for authors, countries of the corresponding author, keywords, and other categories.

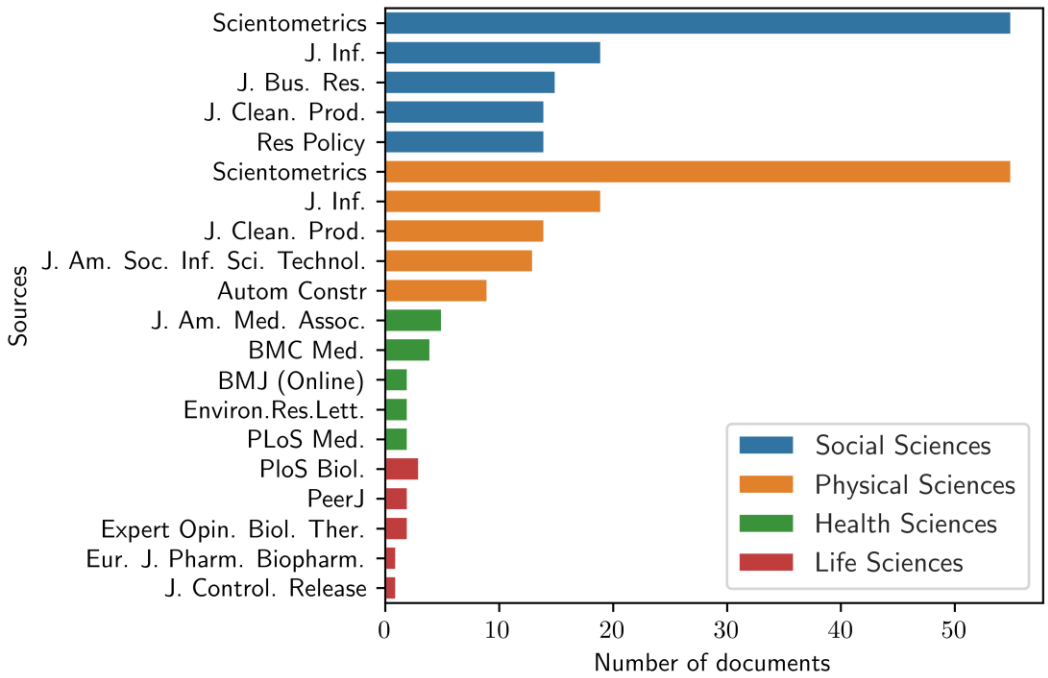


Figure 5. Bar chart from Biblium: top sources contributing to documents in analyzed groups: Social Sciences, Physical Sciences, Health Sciences, and Life Sciences.

Analysis of group overlapping

When analyzing groups, a document can belong to multiple groups (such as in our case, where it belongs to multiple scientific disciplines), meaning that some groups

may overlap. This overlap is implemented in Biblium through various visualizations, including upsetplot, heatmaps, clustermaps, network graphs, and Venn diagrams. Heatmaps and clustermaps depict the total number of overlapping documents or their normalized values, with Jaccard index as the default normalization method, though other indices are available. The overlap can be plotted using a snippet

```
>>> bg.plot_overlapping(kind="venn")
```

that generates a Venn diagram. The final result of this visualization is shown in Figure 6. If the groups are completely disjoint, these visualizations are meaningless, and Biblium does not provide them.

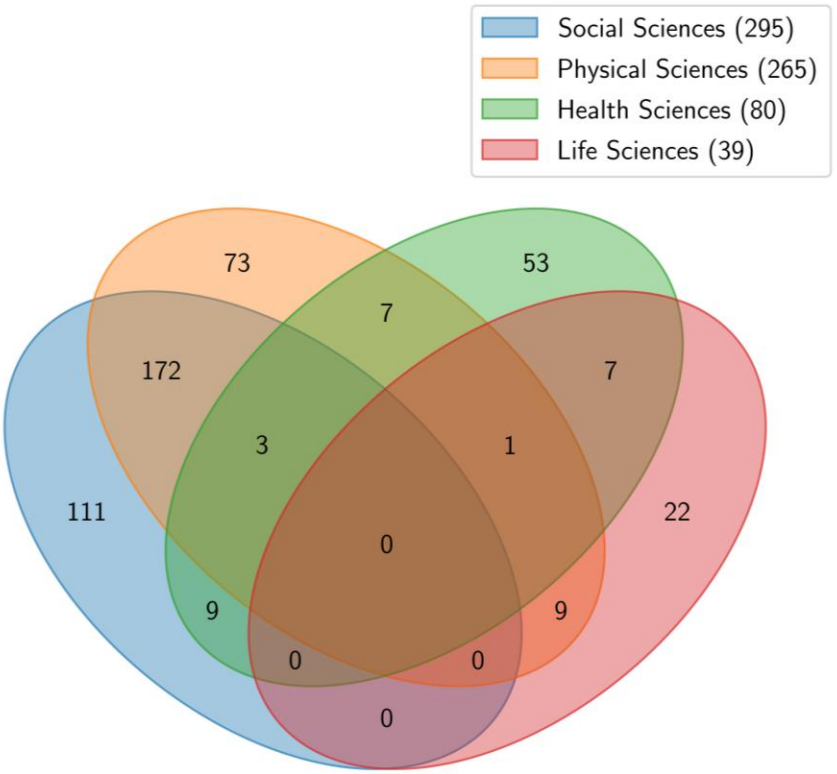


Figure 6. Venn diagram from Biblium: distribution of documents across four scientific disciplines (Social Sciences, Physical Sciences, Health Sciences, and Life Sciences). Overlaps indicate documents categorized under multiple disciplines.

This type of visualization, as shown in Figure 6, effectively captures overlapping and non-overlapping regions for up to four groups. While it is technically possible to represent up to six groups in a Venn diagram, the plot becomes increasingly complex and difficult to interpret as the number of groups exceeds four. In a Venn diagram, the sizes of the groups are included by default in the legend for reference, helping to understand the overall sizes of the groups.

For datasets with more than four groups, a heatmap may be a better choice for visualizing pairwise overlaps, as it is not limited by the number of groups and provides a clearer representation of relationships between categories. However, more complex overlaps involving more than two groups cannot be shown in a heatmap.

Scientific production

The scientific production over time for all groups can be computed in a manner similar to computation in BiblioAnalysis. Here, the resulting plot provides insights into the distribution of published documents across all groups and overall cumulative citations. The script snippet

```
>>> bg.get_production()
>>> bg.plot_production(cut_year=2000)
```

produced the graph shown in Figure 7.

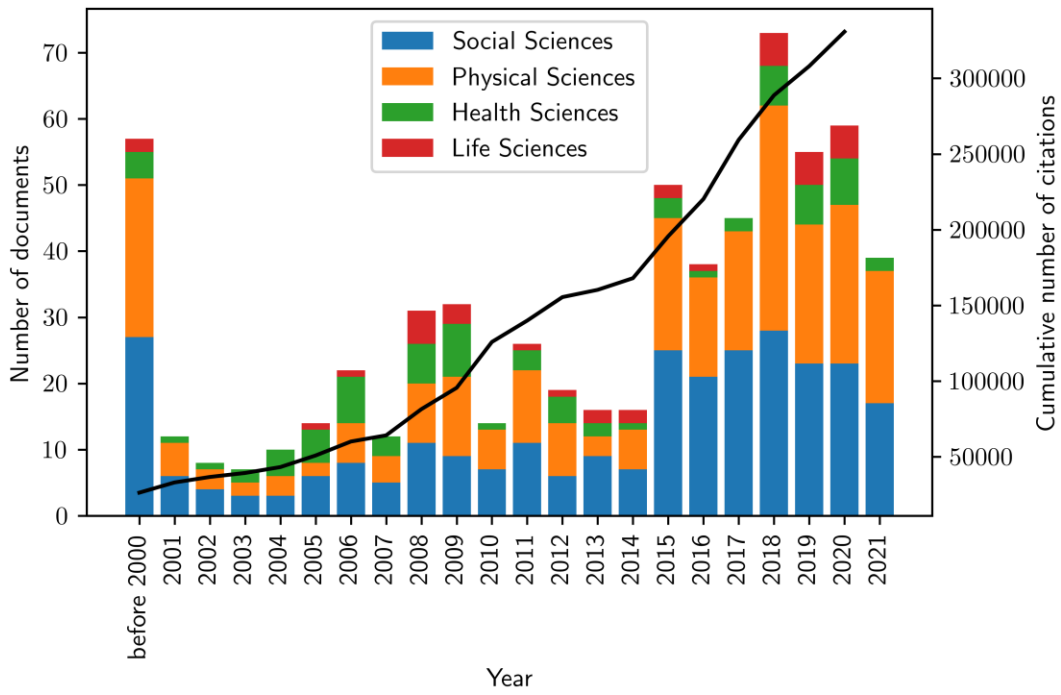


Figure 7. Stacked bar chart from Biblium: annual number of documents categorized by scientific disciplines, alongside cumulative citations over time.

The figure presents a stacked bar chart with a cutoff year, categorizing documents into overlapping disciplinary groups. The black line represents the cumulative number of citations. Since the groups are not pairwise disjoint, the total number of documents in any given year can exceed the actual count due to overlaps, requiring careful interpretation. This visualization provides an aggregated view of the temporal trends in scientific output and its impact.

Measuring Associations

In bibliometric group analysis, various concepts such as sources, authors, and keywords can be associated with divisions into groups, enabling a structured examination of their relationships. The general approach involves computing a 2×2 contingency table for each association, where documents are classified based on whether they belong to a particular group (yes/no) and whether they represent the given concept (yes/no). This results in four possible combinations, forming the basis for statistical analysis. From these tables, multiple measures can be derived to quantify the strength and significance of associations. By default, Biblium computes counts, marginal proportions, the Jaccard index, Yule's Q, and the odds ratio, along with Fisher's exact test to assess statistical significance. While many additional measures are implemented in Biblium, they remain disabled by default.

Let us illustrate the computation of association on a concrete pair (group, keyword). The table 2 shows the relationship between group membership (Physical Sciences or not) and keyword assignment ("industry 4.0" or not). The count a = 10 represents the number of documents in the Physical Sciences group having assigned keyword "industry 4.0," while b = 145 refers to those in the same group but without this keyword. Similarly, c = 0 indicates no documents outside the Physical Sciences group are linked to "industry 4.0," whereas d = 107 represents those outside this group and without the keyword. In total, 262 documents were analyzed, a number lower than the total available (under 500) since those without author keywords were automatically excluded.

Table 2. Contingency table behind calculations from Biblium for illustration: distribution of documents categorized by their membership to "Industry 4.0" and Physical Sciences. This table is a starting point for statistical calculations.

	"industry 4.0"	*"not industry 4.0"*	*Total*
Physical Sciences	a=10	b=145	155
not Physical Sciences	c=0	d=107	107
Total	10	252	262

From this table Jaccard index can be computed using the formula

$$J = \frac{a}{a+b+c} = \frac{10}{10+145+0} = 0.0645$$

and Yule's Q using formula

$$Q = \frac{ad-bc}{ad+bd} = \frac{10*107-145*0}{10*107+145*0} = 1$$

A Yule's Q value of 1 indicates a perfect positive association between the two variables. These measures provide insight into the strength and overlap of the association between group membership and keyword usage. In addition to descriptive measures like Jaccard and Yule's Q, a statistical test such as Fisher's exact test can be performed on this contingency table to assess whether the observed

association is statistically significant. This test is implemented in Biblium by default, providing a rigorous method to evaluate the strength and significance of such associations. Alongside the p-value, Biblium also computes the odds ratio (OR) to quantify the strength of the association. Since multiple hypotheses are tested simultaneously, p-values are adjusted to control for false discovery rates. By default, Biblium applies the Benjamini-Hochberg FDR (Benjamini & Hochberg, 1995) method to ensure reliable statistical inference.
The snippet

```
>>> bg.associate_keywords()
```

computes associations between all groups in the dataset and a selected set of keywords, which by default includes the top 20 keywords across the entire dataset. These associations can be calculated for various units, such as sources, authors, countries of corresponding authors, and more, depending on the analysis focus. For illustration, the table 3 displays only a subset of group-keyword pairs and selected measures. While the analysis generates a range of metrics (including raw counts a, b, c and d from the contingency table), the table highlights key measures such as the Jaccard index, Yule's Q, and the p-value from Fisher's exact test. For illustration, only associations with an unadjusted p-value less than or equal to 0.05 are included in the table 3.

Table 3. Part of Biblium keyword association analysis (computed keywords_assoc_df dataframe): keyword-group pairs with Jaccard index, Yule's Q, and p-values less than 0.05.

group	*keyword*	*Jaccard*	*Yule Q*	*p-value*
Physical Sciences	industry 4.0	0.065	1.000	0.006
Social Sciences	co-citation	0.142	0.698	0.008
Physical Sciences	web of science	0.127	0.596	0.009
Life Sciences	scientometrics	0.129	0.720	0.014
Social Sciences	network analysis	0.077	1.000	0.014
Social Sciences	sustainability	0.087	0.733	0.049
Social Sciences	co-citation analysis	0.082	0.717	0.050

In Biblium, the association of keywords with groups can also be visually explored through word clouds. For each group one word cloud is plotted. In each plot, the size of a word in the word cloud reflects its frequency in the dataset, while the color represents its strength of association with the group, based on a selected measure of association (e.g., Jaccard index, Yule's Q). This visualization provides an intuitive way to interpret both the prevalence and the relevance of keywords within specific groups, enhancing the understanding of the relationships in the dataset.

Other functionalities

In Biblium, bibliometric group analysis is further enhanced by the PredictBiblioGroup class, which is inherited from the BiblioGroupAnalysis class. This implementation integrates data mining prediction methods, where group membership serves as the dependent variable. Predictive modeling in PredictBiblioGroup supports all statistical prediction models available in Scikit-learn (Pedregosa Fabian et al., 2011), along with logistic regression from Statsmodels (Seabold & Perktold, 2010). These models facilitate the classification of bibliometric entities based on various predictive features. To ensure the reliability of predictions, model evaluation is performed using 5-fold cross-validation by default. Performance assessment includes standard classification metrics such as classification accuracy and area under the curve (AUC), among others.

Another class inherited from BiblioGroupAnalysis focuses on the identification of research related to Sustainable Development Goals (SDGs). This method relies on predefined queries from Scopus, where selected keywords and rules determine whether a document (based on its title, abstract, and keywords) is associated with a particular SDG (SDG1–SDG16). This classification approach provides insights into the alignment of scientific output with global sustainability objectives.

This is just one example of a specific application we utilized in our previous research ((Umek et al., 2023), (Umek et al., 2024)). More broadly, this method can be adapted for different thematic groupings. By defining a customized list of keywords relevant to a particular concept, the same approach can be applied to classify documents into other specific research domains.

Conclusion and Future Work

The Biblium project represents an advancement in bibliometric and scientometric analysis within the Python ecosystem, offering a comprehensive and flexible alternative to existing tools. By replicating and expanding upon the core functionalities of Bibliometrix, Biblium introduces key innovations such as (sub)group analysis, predictive modeling, and advanced visualization techniques.

The Biblium project aims for a public release through GitHub, making the source code accessible to the research community and open-source contributors. This step will allow for collaboration, issue tracking, and community-driven improvements. Alongside the GitHub release, Biblium will be packaged as a Python library available via pip, enabling easy installation and integration into research workflows. The pip package will ensure streamlined updates and compatibility with various Python environments.

To enhance Biblium's analytical capabilities, data mining techniques will be incorporated, enabling users to uncover hidden patterns, trends, and relationships within bibliographic datasets. These methods will support text mining, clustering, and classification, facilitating deeper insights into scientific production and citation networks. By integrating advanced algorithms, Biblium will become a powerful tool for researchers looking to analyze large bibliometric datasets with minimal effort.

Biblium will expand its visualization capabilities using Bokeh (Bokeh Development Team, 2018), a powerful library for interactive and high-performance graphics. This will allow users to create dynamic plots, interactive dashboards, and detailed visual representations of bibliographic data. Compared to static plots, Bokeh will enable users to explore their data more intuitively, with zooming, filtering, and hover tools enhancing interpretability. These improvements will be particularly beneficial for analyzing citation networks, keyword co-occurrences, and publication trends.

Future versions of Biblium will incorporate large language models (LLMs) to enhance automated document analysis, clustering, and topic identification. LLMs can be leveraged for the automatic description of document clusters, identification of emerging topics, and a more automated approach to textual and bibliometric analysis. By integrating AI-based techniques, Biblium will provide deeper insights beyond traditional statistical methods, offering richer contextual understanding. One aspect of future research will involve comparing traditional clustering algorithms—such as hierarchical clustering, partitioning methods like k-means, and bibliographic coupling—with AI-driven approaches to evaluate their effectiveness and applicability in structuring scientific knowledge.

A Tkinter-based application (Lundh, 1999) is planned to provide a graphical user interface (GUI) for Biblium, making it accessible to users unfamiliar with coding. The app will include menus, scrollable canvases, and tabbed interfaces for organizing various bibliometric tasks. Users will be able to load, filter, and visualize data through an intuitive interface, with buttons for executing core functions. The Tkinter app will serve as a lightweight, standalone solution for researchers seeking an easy-to-use bibliometric tool.

To further broaden its usability, Biblium will be developed as an add-on for Orange (Demšar et al., 2013), a popular open-source data visualization and machine learning tool. This integration will enable users to apply Biblium's bibliometric and text-mining functionalities within Orange's visual programming environment. Researchers will be able to drag and drop Biblium components into their workflows, combining them with Orange's built-in machine learning and data processing features. This will make Biblium accessible to a wider audience and facilitate exploratory bibliometric analysis without requiring extensive coding knowledge.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency (research programme No. P5-0093 and project No. J5-50183). Finally, in the preparation of this manuscript, the authors utilized ChatGPT, version 4o, developed by OpenAI, for limited and supplementary purposes. Specifically, ChatGPT was employed to assist with checking the grammar, enhancing clarity, and

polishing the language. ChatGPT did not contribute to the intellectual content or scientific insights of the manuscript.

References

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2010). hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*, *82*(2), 391–400. https://doi.org/10.1007/S11192-009-0047-5

Anderson, T. R., Hankin, R. K. S., & Killworth, P. D. (2008). Beyond the Durfee square: Enhancing the h-index to score total publication output. *Scientometrics*, *76*(3), 577–588. https://doi.org/10.1007/S11192-007-2071-2/METRICS

Aria, M., & Cuccurullo, C. (2017a). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/J.JOI.2017.08.007

Aria, M., & Cuccurullo, C. (2017b). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/J.JOI.2017.08.007

Batagelj, V., & Mrvar, A. (2004). *Pajek — Analysis and Visualization of Large Networks*. 77–103. https://doi.org/10.1007/978-3-642-18638-7_4

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc."

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bokeh Development Team. (2018). *Bokeh: Python library for interactive visualization*.

Cabrerizo, F. J., Alonso, S., Herrera-Viedma, E., & Herrera, F. (2010). q2-Index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *Journal of Informetrics*, *4*(1), 23–28. https://doi.org/10.1016/J.JOI.2009.06.005

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/ASI.20317

Cobo, M. J., Lõpez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630. https://doi.org/10.1002/ASI.22688

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, *14*.

Flyamer, I. (2012). adjustText. In *GitHub repository*. GitHub.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). *Exploring Network Structure, Dynamics, and Function using NetworkX*.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van

Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature 2020 585:7825*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Heldens, S., Sclocco, A., Dreuning, H., van Werkhoven, B., Hijma, P., Maassen, J., & van Nieuwpoort, R. V. (2022). litstudy: A Python package for literature reviews. *SoftwareX*, *20*, 101207. https://doi.org/10.1016/J.SOFTX.2022.101207

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/PNAS.0507655102

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Inc., P. T. (2015). *Collaborative data science*. Plotly Technologies Inc.

Jin, B. H., Liang, L. M., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, *52*(6), 855–863. https://doi.org/10.1007/S11434-007-0145-9/METRICS

Laserson, U. (2009). Sqaurify. In *GitHub repository*. GitHub.

Lundh, F. (1999). An introduction to tkinter. *URL: Www. Pythonware. Com/Library/Tkinter/Introduction/Index. Htm*.

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*.

McLevey, J., & McIlroy-Young, R. (2017). Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics*, *11*(1), 176–197. https://doi.org/10.1016/J.JOI.2016.12.005

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional de La Información*, *29*(1), 1699–2407. https://doi.org/10.3145/EPI.2020.ENE.03

Mueller, A. (2010). Word cloud. In *GitHub repository*. GitHub.

Nothman, J. (2023). *UpSetPlot*. https://github.com/jnothman/UpSetPlot

OpenGenderTracking. (2013). https://github.com/OpenGenderTracking/GenderTracker

Pedregosa Fabian, Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay Y, Fré. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Peirson, B. R. E. (2016). *Tethne v0.7. http://diging.github.io/tethne/*. Et Al.

Pereira, V., Pereira Basilio, M., Henrique, C., & Santos, T. (2025). PyBibX-a Python library for bibliometric and scientometric analysis powered with artificial intelligence tools. *Data Technologies and Applications*. https://doi.org/10.1108/DTA-08-2023-0461

Rehurek, R., & Sojka, P. (2011). Gensim--python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Rose, M. E., & Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, *10*, 100263. https://doi.org/10.1016/J.SOFTX.2019.100263

Rossetti, G., Milli, L., & Cazabet, R. (2019). CDLIB: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, *4*(1), 1–26. https://doi.org/10.1007/S41109-019-0165-9/TABLES/5

Ruiz-Rosero, J., Ramirez-Gonzalez, G., & Viveros-Delgado, J. (2019). Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications. *Scientometrics*, *121*(2), 1165–1188. https://doi.org/10.1007/S11192-019-03213-W

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Tctianchi. (2014). Pyvenn. In *GitHub repository*. GitHub.

Team, S. (2009). *Science of Science (Sci2) Tool*. https://sci2.cns.iu.edu

Umek, L., Ravšelj, D., & Aristovnik, A. (2023). Public sector reforms and sustainable development: evidence from bibliometric analys. *IASIA 2023 Conference*.

Umek, L., Takahiro, M., Aristovnik, A., & Ravšelj, D. (2024). Collaborative governance and sustainable development: evidence from bibliometric analysis. *IAS Mombasa Conference 2024*.

Umesh, P. (2012). Image Processing in Python. *CSI Communications*, *23*.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/S11192-009-0146-3

van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*, *111*(2), 1053–1070. https://doi.org/10.1007/S11192-017-2300-7/TABLES/4

Velasquez, J. D. (2023). TechMiner: Analysis of bibliographic datasets using Python. *SoftwareX*, *23*. https://doi.org/10.1016/J.SOFTX.2023.101457

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods 2020 17:3*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

https://doi.org/10.51408/issi2025_149

Book Authors as Self-Promoters on X (Twitter) and Their Information Dissemination Networks

Yajie Wang[1], Haiyan Hou[2], Alesia Zuccala[3]

[1]*yajie.wang@uni-corvinus.hu*
[1]Center for Collective Learning, Corvinus Institute for Advanced Studies (CIAS), Corvinus University of Budapest, 1093 Budapest (Hungary)

[2]*houhaiyan@dlut.edu.cn*
[2]School of Public Administration and policy, Dalian University of Technology, Dalian, 116024 (China)

[3]*a.zuccala@hum.ku.dk*
[3]Department of Communication, University of Copenhagen, Karen Blixens Plads 8, 2300 Copenhagen (Denmark)

Abstract

This is a research-in-progress paper concerning how authors promote their books on X (Twitter), and what follows in terms of an information dissemination network. Our study is based on a sample of books (*n*=2,960) published in 2023 and extracted from Open Alex. While self-promotion is a common and intuitive way to attract the public's attention to one's scholarly accomplishments, little is known about how this leads to further mentions on X (Twitter). From our pilot dataset, we found that 22% of books indexed at OpenAlex exhibit author self-promotion. We then investigated how '*authoritative*' (first tweets) propagate compared to '*connector*' (retweets) and found that this resulted in different types of networks, some we call 'broadcast' networks; others that are 'chain-like'. We also discovered mixed 'broadcast and chain' networks, and it is these that may provide evidence of interdisciplinary research sharing. Further qualitative research is needed to understand the content of this network type.

Introduction

As a scholarly communication channel, Twitter is used by multiple stakeholders, ranging from individual researchers (Holmberg et al., 2014; Ke et al., 2017), libraries (Chu & Du, 2013; Linvill et al., 2012; Veletsianos, 2016; Veletsianos et al., 2017), as well as universities (Kimmons et al., 2017; Linvill et al., 2012). In academia alone, attention has been given to journals (Kortelainen & Katvala, 2012; Ortega, 2017), conference proceedings (McKendrick et al., 2012; Sugimoto et al., 2017), and articles relevant to specific subject areas (Botting et al., 2017; Mahrt et al., 2014). The promotion of articles on social media has been investigated widely (Dixon et al., 2015; Erdt et al., 2017; Fox et al., 2015; Hawkins et al., 2017; Kudlow et al., 2020). Yet, few studies have been carried out pertaining to the dissemination of scholarly books on X (Twitter).

An early study by Thoring (2011) found that the size of a book publisher affects whether or not it will use Twitter for promotional purposes. Snijder (2016) also

discovered that if a monograph is an open access publication, this increases the degree to which it is both tweeted and cited. Further research by Wang and Zuccala (2021) and Wang et al. (2023) have shown that when publishers use Twitter for promotional purposes, their books are more visible, compared to books mentioned by non-publishers.

In this study we investigate book authors who self-promote on X (Twitter) and the information dissemination network resulting from this act. We hypothesize the presence of three types of information dissemination networks (i.e., an isolated network, chain network, or broadcast network) based on the involvement of three types of 'actor' or network nodes: 1) *authoritative*, 2) *connector*, or 3) *propagator* (see **Figure 1**).

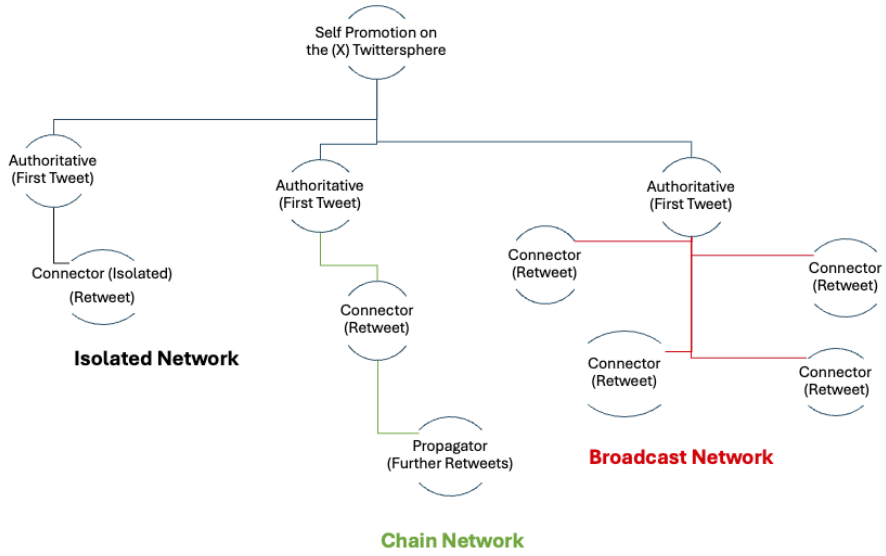


Figure 1. Three types of nodal 'sharing' roles on X (Twitter) and resulting information dissemination networks.

Our motivation for conducting this research relates to the earlier work of Wu et al. (2011), Havakhor et al. (2018), Liang (2018), Wu and Wu (2021), as well as Liu et al. (2023). According to Liu et al. (2023), data used to model how information spreads via social networks, or amongst users on social media, can be both explanatory and predictive.

For example, Havakhor et al., (2018) examined how reputations grow on social media, and found two distinct mechanisms on Twitter: 1) adaptive and 2) objective, each of which corresponded with three knowledge roles: 1) seekers, 2) contributors, and 2) brokers. Although the reputation mechanisms and roles consistently interacted, findings revealed that it was the 'broker' role that 'outperformed' the others. In a

similar vein, Liang (2018) examined patterns of diffusion related to political messages on Twitter, and discovered that a viral diffusion model, in contrast to a broadcast model, increased the likelihood of cross-ideological sharing. Here, the objective is to identify how prevalent it is for authors to 'authoritatively' self-promote their books on Twitter and to examine which type of subsequent dissemination network tends to occur the most.

One reason for mentioning academic work (i.e., in this case books) on a social media platform like (X) Twitter is to ensure that it spreads or reaches as many individuals as possible - i.e., not just 'friends' but also 'friends of friends'. This requires constructing ego networks and examining the nodes to whom the "ego" is directly connected (i.e., 1st-degree ties) plus further ties (i.e., 2nd-degree ties), if any. Ego-networks not only reveal how visible books are in general on (X) Twitter but provide insights into where the presence of *connectors* (1st-degree ties) and *propagator* (2nd-degree ties) might be an indication of interdisciplinary sharing.

Methodology

A dataset of books (n=46,781) published in 2023 (PY=2023) was extracted from OpenAlex on December 3rd, 2024. To determine the X (Twitter) activity associated with these books, we used Altmetrics Explorer at Altmetric.com and relied on each book's DOI or ISBN for retrieval purposes.

Starting with n=46,781 books, we found that a total of n=12,191 books had received mentions on Twitter. However, most of these tweets lacked author information – i.e., we could not determine if it was the author of the book that made the tweet, and the reason for this remains unclear. Our final dataset therefore consisted of n=2,960 books, with authors clearly identified, and where each book had been mentioned at least once on Twitter.

Matching author names to Twitter user-accounts

To identify the authors-as-tweeters, we examined every book for potential matches utilizing a binary approach, like Costas & Mongeon (2020). This procedure involved extracting all the book authors' full names from the OpenAlex records and employing either a "containment-matching" or 'token-matching approach' (Peng et al., 2022). If the names on X (Twitter) consisted of a single-token string – i.e., the author's first name (or last name) had at least 4 characters, it was a 'containment' match; otherwise, the "token-matching" approach meant that the first name or the last name should be matched to the tokens of tweet names (i.e., split by space or underscore).

Classifying the 'ego' network nodes

All the authors-as-tweeters retrieved were classified according to one of four types of network nodes, based on tweet/re-tweet behaviors: 1) the *authoritative* 2) the

connector 4) the *propagator,* and 4) the *isolate*. The *authoritative* is one who possesses an in-degree =0 and out-degree >0. This type is always retweeted by others, but they themselves never retweet. *Connectors* are users with an in-degree > 0 and out-degree > 0. Whilst they only retweet once, they may be further retweeted by others. The *propagators* have an in-degree >1, and an out-degree=0, since they retweet many other user's tweets, but are not retweeted (i.e., unless we include 3rd degree propagators). And finally, *isolates* possess an in-degree =1 and an out-degree =0.

Preliminary Results

Author self-promotion (ASP) on (X) Twitter

A total of *N*=664 of the N=2,960 books from our working dataset (22.4%) could be traced back to an author's Twitter account and identified as being a self-promoter (ASP).

Table 1. Frequency and percentage of tweets and retweets of OpenAlex books.

	Total #books	#Original tweets per book	#Retweets per book	%Original tweets per book	%Retweets per book
Books: ASP	664 (22.4%)	4,360	13,244	24.8%	75.2%
Books: Not ASP	2,296 (77.6%)	7,188	16,619	30.2%	69.8%

Amongst the *n*=664 author self-promoted books, the majority were single authors (87%). The self-promoted books co-authored by two or more authors represented less than 13% of the data in our dataset. Single authors are therefore more inclined to post original tweets about their books on X (Twitter) compared to co-authors (see **Figure 2**).

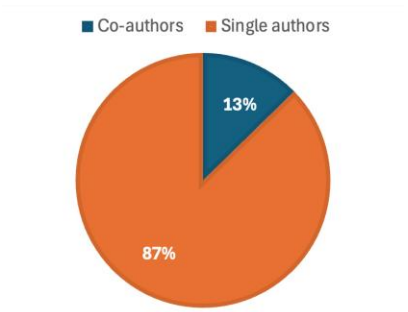


Figure 2. Proportion of single book authors versus co-authors.

Author self-promotion (ASP) roles

We then further categorized the network nodal roles of all the book authors based on their self-promotion efforts (see **Figure 3** and **Figure 4**). We found that single authors primarily took an *authoritative* role (48.4%), followed by a *connector* role (30.9%) in the overall information diffusion process. Co-authors, on the other hand, tended to function primarily as *connectors* (35.4%) or *isolates* (14.1%). This suggests that single authors often undertake self-promotion via original tweets, whereas co-authors are more likely to retweet, or 'connect' the initial tweet of someone else.

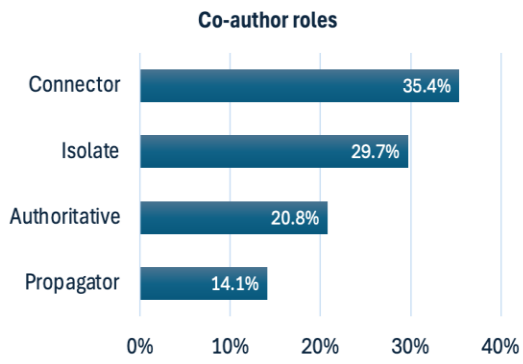


Figure 3. Percentages of the different types of co-author roles in the X (Twitter) dissemination network.

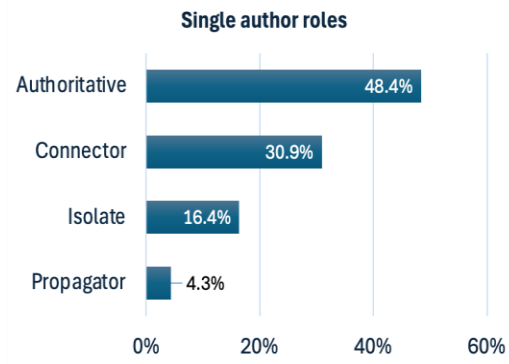


Figure 4. Percentages of the different types of single author roles in the X (Twitter) dissemination network.

Authoritative self-promotion and network types

Here we examine the information dissemination networks for authors who play an 'authoritative' network role (n=97/664; 14.6%) as well as those who play a 'connector' role. **Figures 5** and **Figure 6**, below, present two 'birds' eye views of the networks, constructed using Gephi. Each demonstrates that the prospects for information diffusion are quite different depending on the role that an author plays on X (Twitter). The first Twittersphere (**Figure 5**) is less interconnected than the second (**Figure 6**), therefore authors who self-promote '*authoritatively*' tend to achieve less visibility compared to those who self-promote by 'retweeting', or 'connecting' to another X(Twitter) users' endorsement' (first tweet).

Figure 5 specifically illustrates the presence of 'isolated' networks, as well as small 'chain networks' based on one *authoritative* node, a *connector* tweet (1st degree tie) and a *propagator* tweet (2nd-degree tie) tweet. Here we also see the prevalence of various broadcast networks, where one *authoritative* author node is linked to multiple different *connector* nodes (1st-degree ties). The presence of mixed broadcast and chain networks, shown up close in **Figure 7**, indicates where a content analysis of individual tweets might provide evidence of interdisciplinary information sharing on X (Twitter).

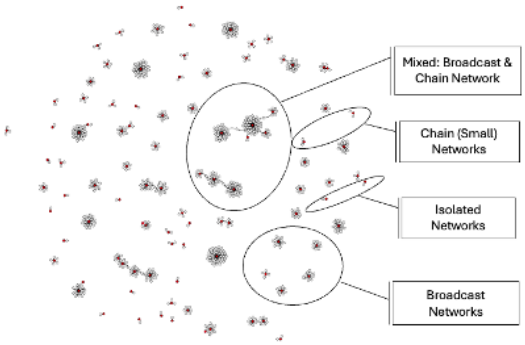


Figure 5. Authors who self-promote their books on X (Twitter) via an *authoritative* role (i.e., first tweets).

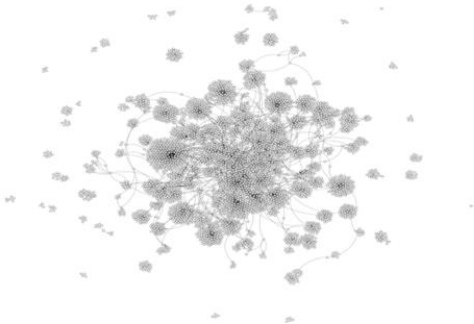


Figure 6. Authors who self-promote their books on X (Twitter) via a *connector* role (retweets).

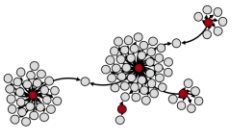


Figure 7. Mixed broadcast and chain networks. Ego-node (red) is self-promoting author.

Conclusions

Our analysis reveals that single-authored books dominate author self-promotion (ASP) efforts on Twitter, with nearly half of the authors acting as authoritative nodes. The hybrid broadcast-chain structures observed in 14.6% of authoritative ego networks suggest latent opportunities for cross-disciplinary engagement. Future research will combine ego networks to identify role shifts from authoritative to connector or propagator over time and whether these shifts affect book visibility. We also plan to expand cross-platform analysis and compare X (Twitter) dissemination patterns with Mastodon/BlueSky to assess the degree of platform dependency in scholarly book promotion.

Acknowledgments

Yajie Wang received the 101086712 LearnData-HORIZON-WIDERA-2022-TALENTS-01 grant financed by the EUROPEAN RESEARCH EXECUTIVE AGENCY (REA) (https://cordis.europa.eu/ project/id/101086712).

References

Botting, N., Dipper, L., & Hilari, K. (2017). The effect of social media promotion on academic article uptake. *Journal of the Association for Information Science and Technology, 68(3),* 795-800.

Chu, S. K.-W., & Du, H. S. (2013). Social networking tools for academic libraries. *Journal of librarianship and information science, 45(1)*, 64-75.

Dixon, A., Fitzgerald, R. T., & Gaillard, F. (2015). Letter by Dixon et al regarding article: "A randomized trial of social media from circulation". *Circulation, 131(13),* e393-e393.

Erdt, M., Aung, H. H., Aw, A. S., Rapple, C., & Theng, Y.-L. (2017). Analysing researchers' outreach efforts and the association with publication metrics: A case study of Kudos. *PloS One, 12(8),* e0183217.

Fox, C. S., Bonaca, M. A., Ryan, J. J., Massaro, J. M., Barry, K., & Loscalzo, J. (2015). A randomized trial of social media from Circulation. *Circulation, 131(1),* 28-33.

Havakhor, T., Soror, A. A., & Sabherwal, R. (2018). Diffusion of knowledge in social media networks: effects of reputation mechanisms and distribution of knowledge roles. *Information Systems Journal, 28(1)*, 104-141.

Hawkins, C. M., Hunter, M., Kolenic, G. E., & Carlos, R. C. (2017). Social media and peer-reviewed medical journal readership: a randomized prospective controlled trial. *Journal of the American College of Radiology, 14(5),* 596-602.

Holmberg, K., Bowman, T. D., Haustein, S., & Peters, I. (2014). Astrophysicists' conversational connections on Twitter. *PloS one, 9(8),* e106086.

Ke, Q., Ahn, Y.-Y., & Sugimoto, C. R. (2017). A systematic identification and analysis of scientists on Twitter. PloS one, 12(4), e0175368.

Kimmons, R., Veletsianos, G., & Woodward, S. (2017). Institutional uses of Twitter in US higher education. *Innovative Higher Education, 42,* 97-111.

Kortelainen, T., & Katvala, M. (2012). "Everything is plentiful—Except attention". Attention data of scientific journals on social web tools. *Journal of Informetrics, 6(4),* 661-668.

Kudlow, P., Bissky Dziadyk, D., Rutledge, A., Shachak, A., & Eysenbach, G. (2020). The citation advantage of promoted articles in a cross-publisher distribution platform: a 12-month randomized controlled trial. *Journal of the Association for Information Science and Technology, 71(10),* 1257-1274.

Liang, H. (2018). Broadcast versus viral spreading: The structure of diffusion cascades and selective sharing on social media. *Journal of Communication, 68(3),* 525-546.

Linvill, D. L., McGee, S. E., & Hicks, L. K. (2012). Colleges and universities' use of Twitter: A content analysis. *Public Relations Review, 38(4),* 636-638.

Liu, Y., Zhang, P., Shi, L., & Gong, J. (2023). A Survey of Information Dissemination Model, Datasets, and Insight. *Mathematics, 11(17),* 3707.

Mahrt, M., Weller, K., & Peters, I. (2014). Twitter in scholarly communication. *Twitter and Society,* 89, 399-410.

McKendrick, D. R., Cumming, G. P., & Lee, A. J. (2012). Increased use of Twitter at a medical conference: a report and a review of the educational opportunities. *Journal of Medical Internet Research, 14(6)*, e176.

Ortega, J. L. (2017). The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations). *Aslib Journal of Information Management, 69(6),* 674-687.

Snijder, R. (2016). Revisiting an open access monograph experiment: measuring citations and tweets 5 years later. *Scientometrics, 109(3),* 1855-1875.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology, 68(9),* 2037-2062.

Thoring, A. (2011). Corporate tweeting: Analysing the use of Twitter as a marketing tool by UK trade publishers. *Publishing Research Quarterly, 27*, 141-158.

Veletsianos, G. (2016). Social media in academia: Networked scholars. Routledge.

Veletsianos, G., Kimmons, R., Shaw, A., Pasquini, L., & Woodward, S. (2017). Selective openness, branding, broadcasting, and promotion: Twitter use in Canada's public universities. *EducationalMedia International, 54(1)*, 1-19.

Wang, Y., & Zuccala, A. (2021). Scholarly book publishers as publicity agents for SSH titles on Twitter. *Scientometrics, 126(6),* 4817-4840.

Wang, Y., Zuccala, A., Hou, H., & Hu, Z. (2023). Corrigendum to ['To tweet or not to tweet?' A study of the use of Twitter by scholarly book publishers in Social Sciences and Humanities', *Journal of Informetrics, 17(2),* 101351. https://doi.org/https://doi.org/10.1016/j.joi.2022.101351

Wu, B., & Wu, C. (2021). Research on the mechanism of knowledge diffusion in the MOOC learning forum using ERGMs. *Computers & Education, 173,* 104295.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. Proceedings of the 20th international conference on World wide web.

https://doi.org/10.51408/issi2025_126

Catalytic Effect of Open Data Platforms: The Case of the Global Biodiversity Information Facility (GBIF)

Honami Numajiri[1], Michio Oguro[2], Takayuki Hayashi[3]

[1] *doc22053@grips.ac.jp*, [3] *ta-hayashi@grips.ac.jp*
National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku, Tokyo (Japan)

[2] *mog@ffpri.affrc.go.jp*
Forestry and Forest Products Research Institute, 1 Matsunosato, Tsukuba, Ibaraki (Japan)

Abstract

In recent years, data platforms have become essential to scientific research, and the Global Biodiversity Information Facility (GBIF) has emerged as a key infrastructure. However, how such platforms influence research trajectories remains poorly understood. This study examined GBIF's impact on researchers' topic selection by analysing papers before and after GBIF use and comparing them with non-GBIF users. Analysis of the research papers identified 20 distinct topics, revealing shifts in the research focus. GBIF users showed transitions from studies focused on ecosystem processes to research connecting ecosystem processes with plant-level characteristics, while maintaining unique patterns in forest biodiversity conservation. Furthermore, comparative analysis with non-GBIF users demonstrated that GBIF users exhibited unique topic transition patterns, particularly in areas such as plant-host interactions and habitat management, which indicate researchers' growing interest in this field, potentially aligning with increasing policy attention to biosecurity issues. These findings demonstrate GBIF's role as a catalyst in biodiversity science, actively shaping research directions rather than merely serving as a data repository. While highlighting GBIF's significant impact on scientific research, this study also identifies areas that require enhanced data coverage and accessibility to maximise the platform's scientific and societal contribution.

Introduction: The Evolution and Impact of Open Data in Scientific Research

The landscape of scientific research has been significantly transformed through open data policies, reshaping research practices, and knowledge dissemination. Government agencies, research institutions, and funding organizations actively promote research data release as a driver of scientific progress (Hrynaszkiewicz & Cadwallader, 2021). UNESCO defines Open Science as that which "enables free access to and reuse of scientific knowledge, thereby supporting collaboration and knowledge sharing" (UNESCO, 2021). Within Open Science, open research data is freely accessible, reusable, and properly documented data for scientific inquiry (OECD, 2015). Previous research has shown that open data policies increase citation rates, enhance collaboration opportunities, and improve research efficiency (Piwowar et al., 2007; Tenopir et al., 2011). Their significance lies in generating novel research through data reuse by diverse researchers (Borgman, 2015), particularly relevant as data-driven approaches become prevalent (Numajiri & Hayashi, 2024).

Two main types of Open Data exist: researcher-published data and Open Government Data (OGD). Researcher-published data comes from specific studies, providing insights into particular phenomena and enabling replication studies (Zuiderwijk et al., 2020). OGD, made available by government agencies, offers long-term datasets valuable for studying trends like climate change or population dynamics (Wirtz et al., 2022; Zuiderwijk & Janssen, 2014). Infrastructure challenges, particularly data fragmentation across institutions, remain a key barrier to efficient research data utilization (Quarati et al., 2021). While organizations like U.S. Geological Survey and various countries are developing centralized platforms to address this (Ojo et al., 2016), their effectiveness remains uncertain. Even when governments make data available, limited openness and continued fragmentation across agencies create barriers to data discovery and reuse (Wang & Shepherd, 2020).

GBIF: A Global Infrastructure for Biodiversity Open Data & Research question

The Global Biodiversity Information Facility (GBIF) [1] is a significant open data infrastructure in biodiversity research, providing worldwide access to biodiversity data. Before GBIF's 2001 establishment, accessing biodiversity information was challenging due to technical difficulties in data storage, insufficient training, and institutional sharing complications (Jones et al., 2006; Quarati et al., 2021). GBIF was designed to enable novel research through a single web interface (Telenius, 2011). Since 2012, it has seen explosive growth in scientific usage and currently hosts over 1.9 billion species occurrence records from thousands of institutions (GBIF, 2021). This extensive dataset supports applications from climate change assessment to species distribution modelling (Heberling et al., 2021). GBIF has enhanced its infrastructure through improved data management technologies and computational capabilities. Analysis of 4,000+ GBIF-enabled studies (2003-2019) shows species distribution modeling as a dominant application, though research applications have diversified (Heberling et al., 2021).

This study proposes viewing open data infrastructures like GBIF as enabling platforms in scientific research that facilitate novel combinations of research fields. Unlike temporary research funding, GBIF provides persistent infrastructure that enables researchers to integrate diverse data sources and explore new research directions. This catalytic effect leads researchers to discover and pursue new research topics they couldn't have explored before. The research question in this study is: **How does the adoption of GBIF data influence researchers' research topics?** The studies propose the following hypotheses: **GBIF data adoption leads to greater changes in research topics compared to non-users of GBIF data.** Through examining longitudinal changes in research content before and after researchers' initial use of GBIF data, this study investigates how open data infrastructure shapes the trajectory of scientific inquiry.

[1] https://www.gbif.org/

Data and Methodology

From the 'Peer-reviewed papers using data' section of the GBIF website, 10,915 papers with GBIF data DOIs were collected. Scopus search yielded 7,381 papers, from which 29,204 unique author IDs were extracted. All papers by these authors (762,123 papers) were collected through Scopus API. A control group was established from the top 20 journals that published most papers citing GBIF, out of 138 journals with 10+ papers citing GBIF. The top 20 journals yielded 46,811 non-GBIF authors who published a total of 1,156,791 papers. Changes in research topics were analyzed using five-year windows before and after authors' first GBIF use (reference year). Final analysis included 7,171 GBIF authors and 28,022 Non-GBIF authors with 3+ years of papers before/after reference years. The reference year was excluded to isolate GBIF's direct influence on subsequent research. For example, for a 2015 first use, analysis covered 2010-2014 and 2016-2020. For Non-GBIF Group authors, having no GBIF usage, each year from 2010 to 2020 was set as a potential reference year, creating 11 separate datasets per author.

For each author, titles and abstracts of their papers were combined into pre- and post-GBIF document sets. Titles and abstracts were preprocessed using NLTK and scikit-learn, converted to embeddings using Sentence Transformer, and clustered using K-means. Topic changes were measured using cosine distance between embedding vectors (0=identical, 1=dissimilar). To address sample size disparity (Non-GBIF Group having 11 datasets per author), bootstrap sampling matched case numbers between the groups by measuring GBIF Group author counts per reference year. Non-GBIF cases were randomly sampled 100 times to equalize author numbers for each reference year, enabling controlled analysis of topic changes.

Results1: Identification and Analysis of Research Topics

Based on evaluation metrics (silhouette=0.033, Calinski-Harabasz=7209.33, Davies-Bouldin=3.40) and elbow method analysis, the documents were clustered into 20 topics, striking a balance between interpretability and computational efficiency. Table 1 presents these 20 research topics with labels.

Table 1. Twenty research topics.

Topic	Topic label	Topic	Topic label
1	Forest carbon cycling models	11	Taxonomy & phylogeny
2	Conservation of avian habitats	12	Ecology of fish and fishery
3	Carbon stock in forests and biodiversity	13	Genomics of plant traits
4	Modelling	14	Limnology
5	Plant physiology of growth	15	Material cycling in soil
6	Plant diversity and traits in ecosystems	16	Effects of ecosystem changes on biodiversity and ecosystem services
7	Conservation of forest biodiversity and species	17	Forest pest management
8	Disease control	18	Water cycle model of forests
9	Population genetics, genetic diversity	19	Community of soil fungi and microbes
10	Conservation and management of marine ecosystem against climate change	20	Description of novel species

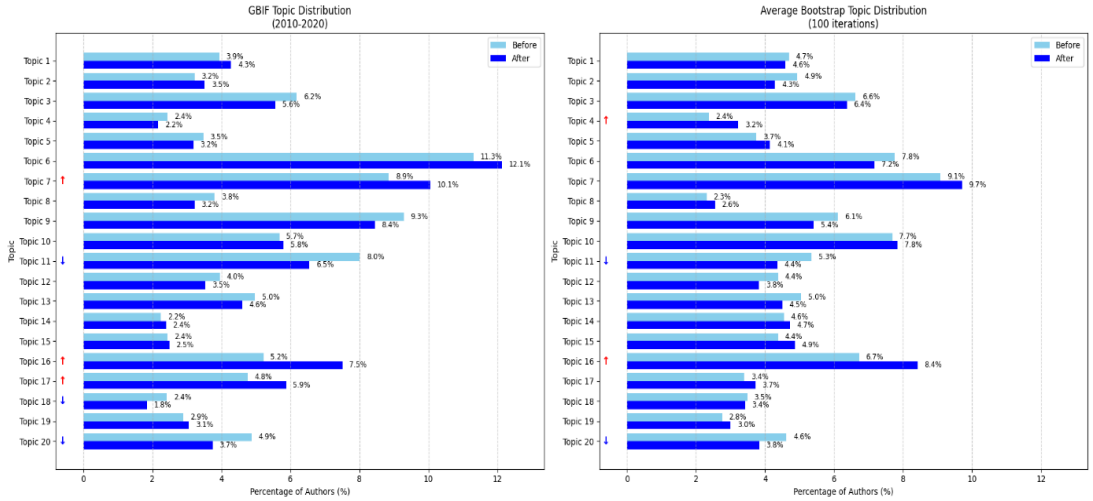


Figure 1 Topic distribution before and after GBIF use in GBIF and non-GBIF Groups (2010-2020).

Left: Topic distribution of GBIF group Right: Average topic distribution of bootstrap samples from non-GBIF group (100 iterations)

Note: The bars show the percentage of authors in each topic before (light blue) and after (dark blue) the reference year. For the GBIF group, the reference year corresponds to the year of their first publication utilizing GBIF data, while for the non-GBIF authors reference year were randomly sampled to match the number of GBIF authors who first used GBIF data in that year.

Comparing topic proportions before/after the reference year revealed key patterns. As shown in Figure , analysis of topic distributions revealed distinct research patterns between GBIF and non-GBIF groups. GBIF users maintained strong engagement in plant-focused research, with Topic 6 (Plant diversity and traits in ecosystems) showing the highest proportion (before: 11.3%, after: 12.1%). They also significantly increased their focus on forest biodiversity conservation (Topic 7: 8.9% to 10.1%, $p < 0.05$) and pest management (Topic 17). Both groups showed increased engagement in ecosystem services research (Topic 16: GBIF 5.2% → 7.5%, non-GBIF 6.7% → 8.4%, $p < 0.001$) and decreased focus on taxonomic studies (Topics 11 and 20). Significant differences in topic composition existed between groups both before ($\chi^2 = 523.400$, $p < 0.01$) and after ($\chi^2 = 626.012$, $p < 0.01$) the reference year. GBIF users showed consistently higher engagement in Topic 9 (Population genetics, genetic diversity), and lower in Topic 14 (Limnology) and 3 (Carbon stock in forests and biodiversity). Unique to the GBIF group were significant decrease in Topic 18 (Water cycle model of forests) ($p < 0.05$).

Results2: Transition Patterns

Given the differences in topic proportions between GBIF and non-GBIF groups described above, Figure 2 illustrates the differences in transition probabilities between topics, comparing GBIF and non-GBIF users. The analysis revealed several distinct patterns in research topic transitions. The most notable pattern centered on

Plant diversity and traits in ecosystems (Topic 6), where the GBIF group showed significant transitions from multiple topics. These included transitions from Topic 1: Forest carbon cycling models (GBIF group 17.1% vs. non-GBIF group 6.7%), from Topic 3: Carbon stock in forests and biodiversity (18.0% vs. 9.3%), and Topic 15: Material cycling in soil (17.3% vs. 11.3%). These transitions show that GBIF users shifted their research focus from forest-level studies to studies incorporating plant functional characteristics.

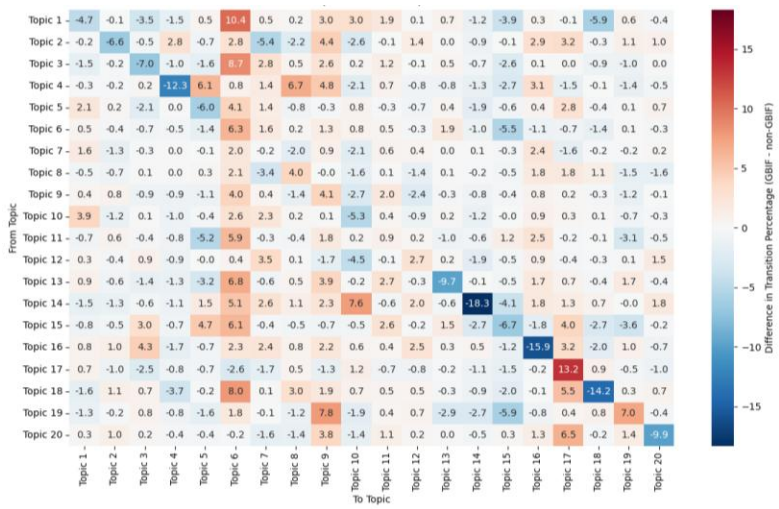


Figure 2 Comparison of topic transition patterns.

Note: Heatmap shows differences in transition probabilities between groups (percentage points). Red cells indicate higher probabilities in GBIF group, blue cells indicate higher in non-GBIF group. Values calculated as GBIF minus non-GBIF probabilities.

Results3: Quantitative Analysis of Topic Transition Distances

An analysis of topic transitions investigated whether GBIF users show greater transition distances between research topics compared to non-GBIF users. As shown in Figure 3, comparing cosine distances between the GBIF dataset (2010-2020, n=5,208, median=0.134, mean=0.135, SD=0.067) and 100 bootstrap samples (95% CI [0.122, 0.130]) revealed that GBIF users' transition patterns exceeded random expectations significantly ($p < 0.01$). The relatively small range of cosine distances (0.1-0.15) suggests that transitions represent expansions within related research domains rather than shifts to entirely different topics.

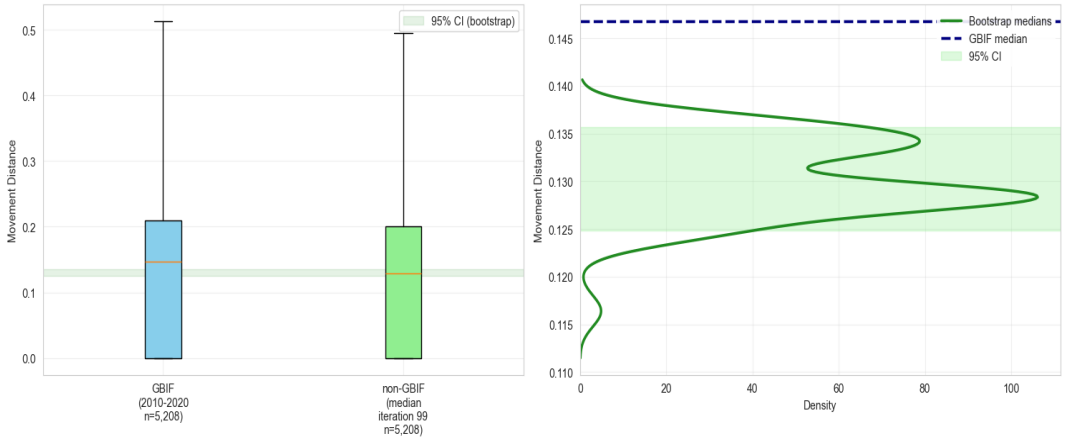


Figure 3 Comparison of Topic Transition Distances between GBIF Data and Bootstrap Samples.

Note: Left panel shows boxplots comparing GBIF and bootstrap samples distributions of non-GBIF. Right panel shows density plots of bootstrap medians, with GBIF median (dashed line) lying outside the bootstrap distribution's main density range (green shaded area).

Discussion and Conclusion

The study highlighted GBIF's impact on research trajectories through analysis of 20 distinct topics. The analysis suggests a possible relationship between GBIF use and researchers' transitions between ecosystem-level processes and plant-level characteristics. As noted by Mandeville et al. (2021), biodiversity platforms like GBIF enhance dataset accessibility, enabling researchers to bridge gaps between research domains. The finding that both GBIF and non-GBIF groups increased their engagement in ecosystem changes and biodiversity services research (Topic 16), while GBIF users showed distinct patterns in forest biodiversity conservation (Topic 7), suggests that while overall research trends may reflect broader scientific interests in the field, access to GBIF data might enable different approaches to biodiversity research.

Notably, the observed transition of GBIF users toward Topic 17 (Forest pest management) might indicate researchers' growing interest in this field. The observed transition patterns toward Topic 17 (Forest pest management) require careful interpretation, as the data suggests transitions from seemingly unrelated research areas such as species description (Topic 20), forest water cycles (Topic 18), and soil material cycling (Topic 15). These unexpected patterns indicate the need for more detailed investigation of how researchers actually shift between research topics.

The significantly higher transition distances in the GBIF group compared to non-GBIF users indicate that GBIF data facilitates broader research exploration, though within related domains as shown by the moderate range of cosine distances (0.1-0.15). This accessibility fosters diverse knowledge integration, leading to thematic diversification (Khan et al., 2021).

This empirical evidence revealed different patterns of research topic transitions between GBIF users and non-users, though the mechanisms and implications of these differences require further investigation.

Reference

Borgman, C. L. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. Big Data, Little Data, No Data. https://doi.org/10.7551/MITPRESS/9963.001.0001

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences of the United States of America, 118(6), e2018093118. https://doi.org/10.1073/pnas.2018093118

Numajiri, H., & Hayashi, T. (2024). Analysis on open data as a foundation for data-driven research. Scientometrics, 1–18. https://doi.org/10.1007/S11192-024-04956-X

Ojo, A., Porwol, L., Waqar, M., Stasiewicz, A., Osagie, E., Hogan, M., Harney, O., & Zeleti, F. A. (2016). Realizing the innovation potentials from open data: Stakeholders' perspectives on the desired affordances of open data environment. IFIP Advances in Information and Communication Technology, 480, 48–59. https://doi.org/10.1007/978-3-319-45390-3_5

Telenius, A. (2011). Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany, 29(3), 378–381. https://doi.org/10.1111/J.1756-1051.2011.01167.X

Vicent Civera, A., Baptista, P., Chatzivassiliou, E., Cubero, J., Cunniffe, N., et al. (2024). Commodity risk assessment of *Prunus cerasus* × *Prunus canescens* hybrid plants from Ukraine. EFSA Panel on Plant Health. *EFSA Journal*, 22 November 2024. https://doi.org/10.2903/j.efsa.2024.9089

https://doi.org/10.51408/issi2025_133

Conceptualizing Metascience Observatories

Emanuel Kulczycki[1], Johann Mouton[2], Didier Torny[3], Sergio Luiz Monteiro Salles Filho[4], Pei-Ying Chen[5], Juan Rogers[6], Noor Jaleel[7], Mathieu Ouimet[8], Kieron Flanagan[9], Aditi Ashok[10], Cassidy R. Sugimoto[11]

[1] *emek@amu.edu.pl*
Scholarly Communication Research Group, Adam Mickiewicz University in Poznań (Poland)

[2] *jm6@sun.ac.za*
Centre for Research on Evaluation, Science and Technology, Stellenbosch University (South Africa)

[3] *didier.torny@cnrs.fr*
Centre de sociologie de l'innovation, CNRS (I3, UMR 9217), Mines Paris, PSL University (France)

[4] *sallesfi@unicamp.br*
Dept. of Science and Technology Policy, Universidade Estadual de Campinas (Brazil)

[5] *peiychen@iu.edu*
Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington (USA)

[6]*jdrogers@gatech.edu*, [7] *njaleel3@gatech.edu*, [10]*ashokaditi25@gmail.com*, [11]*sugimoto@gatech.edu*
School of Public Policy, Georgia Institute of Technology (USA)

[8] *mathieu.ouimet@pol.ulaval.ca*
Department of Political Science, Université Laval (Canada)

[9]*kieron.flanagan@manchester.ac.uk*,
Manchester Institute of Innovation Research, Alliance Manchester Business School, University of Manchester (UK)

Abstract

Science can serve as a powerful source to inform decision-making at the national and international levels. It can also be a source of reflective information: that is, to provide decision-makers with information about the science and technology (S&T) ecosystem. Scientific information about science—i.e., metascience—can provide decision makers and their advisers with evidence needed to direct research activities, allocate resources, and build collaborative relationships (soft power diplomacy). Several institutions around the world are dedicated to the observation of science—i.e., metascience observatories. However, these vary significantly in scope and function, and little is known about the degree to which they directly inform diplomatic decision-making. Therefore, the goal of this research is first to provide an empirical basis for conceptualizing metascience observatories. Through this work, we can clearly delineate metascience observatories from other types of institutions. The generated registry of metascience observatories will then serve as a platform for understanding the role of metascience in diplomacy.

Introduction

Science, as a social institution (Thorpe, 2013), can have a profound influence in politics and social imaginations (Ezrahi, 2012), directly affecting the political

landscape. Science diplomacy, therefore, can also be driven by scientists as agents who apply their expertise to global challenges in ways that go beyond government-directed diplomacy. Scientists assume an expert role in society and can activate this role within the context of diplomatic relations (Weisskopf, 1969). As noted by Jasanoff (2009): "the very virtues that make democracy work are also those that make science work: a commitment to reason and transparency, an openness to critical scrutiny, a skepticism toward claims that too neatly support reigning values, a willingness to listen to countervailing opinions, a readiness to admit uncertainty and ignorance, and a respect for evidence gathered according to the sanctioned best practices of the moment". Despite this strong connection, there is limited evidence on the formal ways in which she is translated into the diplomatic process.

Science diplomacy has been classified into three main areas (AAAS, 2010): *diplomacy for science* (use of diplomatic action to facilitate science), *science for diplomacy* (use of science to advance diplomatic objectives), and *science in diplomacy* (support of diplomatic processes with scientific evidence). All of these, and particularly the latter two, require a strong evidence base drawn from monitoring the scientific system[1]. Monitoring science—that is, formalized observations of the knowledge ecosystem—is referred to in contemporary parlance as *metascience* (a term that was previously used in a different context; other current labels include 'science of science' and 'research on research') and several organizations dedicated to this activity operate around the globe. Metascience observatories should not be confused with *scientific* research infrastructures which gather scientific data on a specific topic (e.g., CERN). Rather, metascience observatories study how science operates within national and international ecosystems, how funds are allocated to research organizations, what knowledge is produced, and in what form it is communicated. In this way, metascience observatories can inform science diplomacy strategies targeted at facilitating access to national and international research capabilities and data, promoting and attracting talent, as well influencing public opinion, and political and economic leaders at national and international levels (Flink & Schreiterer, 2010).

The goal of this research, therefore, is to conceptualize metascience observatories, with the objective of creating a codified registry of metascience observatories across the globe. For this project we ask: *what are the necessary components of a metascience observatory*? Understanding how metascience observatories operate can provide valuable insights into the dynamics of national and international research systems, allowing for more informed diplomatic strategies. In particular, the evidence produced by metascience observatories can enhance the capacity of science diplomacy to address global challenges, and support evidence-informed policymaking.

[1] In using the term "science", we emphasize that it should be understood to encompass all forms and sectors of knowledge creation (including social sciences, arts, and humanities), evoking the notion of German *Wissenschaft* or Latin *scientia*. In metascience, however, this is restricted to the formal manifestation of this knowledge (e.g., through published and indexed journal articles).

Candidate identification

We took an iteratively inductive and deductive approach to conceptualizing metascience observatories. The highly interdisciplinary and geographically diverse team began with discussions on what constituted metascience observatories within their known spaces and across history. This created an initial set of key criteria. Using these criteria, the team began the generation of a list of "seed candidates". From this, we utilized snowball sampling, by expanding to a larger group of experts who provided additional candidates for investigation. The initial list was highly skewed towards geographies to which our experts were proximal. Therefore, we generated an inclusive list of all countries (including non-recognized states and territories for broadest coverage). We then conducted searches for each of these countries using the name of the country + terms such as "metascience", "research evaluation", "research council", "sci tech policy", and "science diplomacy." This generated a list of 209 candidates for investigation.

Codebook generation and justification

Using these 209 candidates as cases for discussion, we refined the inclusion criteria into a codified codebook, with sequential elimination. That is, an affirmative answer must be received for all questions to be considered a metascience observatory. The absence of a single criterion warrant exclusion. Four main categories, with nine subcategories were generated:

1) **PURPOSE**. Metascience observatories are dedicated to the study of the science and technology (S&T) system.
   a) Are observations of the S&T system one of the primary functions of the organization?

2) **FORM**. Metascience observatories are formal organizations.
   a) Does it have more than one individual in the organization?
   b) Does it have a division of labor within the organization?
   c) Does it have rules of membership (which dictates the association of products with the organization)?

3) **FUNCTION**. Metascience observatories collect, analyze, and maintain data about the science and technology (S&T) system.
   a) Are the data about science and technology (as opposed to e.g., scientific data)?
   b) Is there evidence of data analysis and interpretation of these data by the organization?
   c) Are data maintained by the organization?

4) **DISSEMINATION**. Products of a metascience observatory are disseminated consistently to a broad audience.
   a) Are the products (i.e., data analysis and interpretation) of the metascience observatory consistently disseminated as an integral part of its mission?

b) Are products of the observatory accessible to the public?

PURPOSE (1) was identified as the first criterion, as the organization should have, as one of its primary objectives, the study of the science and technology (S&T) system. This was functionally a binary distinction with only one inclusion question **(1a)**. Several organizations, as we will see, are dedicated to science and technology, but do not observe this system. Furthermore, some have observation as a small part of their portfolio, but this is not their primary activity. Operationalizing this criterion typically took the form of reviewing the mission statement for these organizations: if the mission did not articulate observation, it was likely not a primary activity.

The **FORM (2)** that an organization took was also a critical component. To serve as an observatory, we argued that the organization must have a degree of formalization and could not be a single investigator or collaborative platform without governance. The true contrast is a contract specifying a deliverable, in this case, metascience analysis and reporting, even when the contract has hierarchical elements in it, such as standard operating procedures, authority systems, among other possibilities (Williamson 1975; Stinchcombe 1990). We consider two major components for a bureaucracy: (a) specialization and (b) rules of membership. As Durkheim (1893) noted, as specialization increases, rules and norms become essential for maintaining coherence. Likewise, Weber (1922) emphasized that bureaucratic administration relies on knowledgeable structured expertise to guide decision-making. This reinforces the necessity of formal governance and institutional membership, without clear rules, the observatory would lack the structure needed for sustained operations and legitimacy. We operationalize specialization by asking whether there is more than one individual **(2a)** and whether those individuals have **(2b)** clearly defined roles and responsibilities. This distinguishes collectives, such as teams under contract, where individuals use the data, but do not formally affiliate to the institution in the products that they create with the use of these data **(2c)**. In contemporary sociology of organizations, these three components would be summarized in the labor contract, fiduciary relations, standard operating procedures, legal status and internal performance systems that constitute the organization as a hierarchy or bureaucracy (Stinchcombe op. cit.).

Although the function might be implied by 1a, we found that while several organizations stated as their mission to observe the S&T system, they did not produce results that provided evidence of this operation. Therefore, **FUNCTION (3)** is concerned with identifying that the organization collects, analyzes, and maintains data about S&T. The first criterion is **(3a)** whether the data are *about* science and technology (i.e., metascience), as opposed to scientific data. This is what fundamentally distinguishes a *scientific* observatory (e.g., an astronomical observatory) from a *metascientific* observatory. An observatory cannot merely collect these data, but it must also add value to the data through analysis and interpretation. Therefore, we look for *evidence* for such analysis **(3b).** This rule out those who may analyze the data, but do not make this analysis or interpretation available. Finally, we seek to look at those institutions who do not merely utilize third-party sources, but collect and **(3c)** *maintain* data themselves.

Finally, we consider a core component of a metascience observatory to be in the **DISSEMINATION (4)** of their data and analysis to the general public. These criteria draw from previous research (focused on the Latin American context) (Macedo & Maricato, 2022) which state the observatories of science and technology "have and make available indicators and/or statistics…indicate the sources of information for these indicators…[and] have services on an *online portal*." We ask first whether the data are disseminated on a regular basis **(4a)**; that is, not merely ad hoc publications as might be produced by a research lab. This requires that dissemination is a codified part of the mission of the organization. The role of the observatory must have consistency over time and in its domain that cannot be reduced to research projects with changing specifications (Macaulay 1963; Scherer 1964). Secondly, we ask whether data are made accessible to the public **(4b).** This removes organizations who only provide data privately to clients.

Conceptualization

Using the criteria as our guide, we can deduce the following definition for a metascience observatory: **"*A metascience observatory is a formal organization dedicated to the collection, analysis, and maintenance of data about the science and technology (S&T) system which disseminates results consistently to a broad audience*."**

Application of codebook

Given this conceptualization and operationalization, we returned to the list of 209 candidate metascience observatories and applied the codebook, with sequential elimination, that is, we examined the inclusion criteria in order and eliminated a candidate as soon as it failed to meet one of the criteria. That point of exclusion was documented and provided below. Five coders initially looked at candidates and discussed points of disagreement. From this initial conversation, four of the five coders examined a larger number of candidates, again resolving disagreements as they were identified. Finally, two coders examined all 209 independently and then resolved conflicts post-hoc, with engagement from the larger group of initial coders. The analysis was conducted unobtrusively, focusing on material found on the Internet. For a few cases, the websites did not load properly at the time of analysis—these were documented as "Technical" difficulties. The number of candidate metascience observatories excluded at each point is summarized in Table 1.

Table 1. Number of candidate metascience observatories excluded at each criterion.

Reason for exclusion	*Number*	*Percentage*
1a: dedication to observation	92	44%
2a: more than one individual	–	–
2b: division of labor	7	3%
2c: rules of membership	1	<1%
3a: metascience	24	11%
3b: analysis and interpretation	25	12%
3c: data maintenance	3	1%
4a: consistent dissemination	7	3%
4b: public dissemination	2	<1%
Technical difficulties	9	4%

The modal exclusion was (1a): 44% of the candidates (n=92) were excluded because they did not have the *observation* of the S&T system as one of their primary goals. This included several academies of science, research councils and private funders, ministries and governmental organizations, professional organizations, and scientific research centers. All of these organizations had missions to support or promote science, but not necessary to observe it.

Seven candidates (3%) failed to meet (2b). These were either small research labs or research consortia, where there was no clear division of labor or specialization within the organization. One candidate was excluded at (2c)—it had a division of labor, however, no publications affiliated with the organization, demonstrating that it did not have an affiliation property.

Twenty-four candidates (11%) were excluded due to the fact that they did not collect or analyze data *about* science and technology (3a). That is, while they had a primary objective of observing science, they did not collect, analyze, or maintain *metascientific* data. Similar to (1a) this category included several academies of science, research councils, and national centers for science and technology; however, those excluded at this stage had stated monitoring of science in their mission but failed to conduct metascience research.

Twenty-five candidates (12%) were excluded for the lack of analysis and interpretation added to the data (3b). The most common type of organization in this category were open science monitors and dashboards. In addition, there were national indices of researchers, corporate databases, and some ministries and other national organizations which collect data, but do not analyze or provide interpretation to the data.

Three candidates were excluded for the lack of evidence of data maintenance (3c). These were largely highly developed research centers, which analyzed third-party data, but did not collect or maintain data themselves.

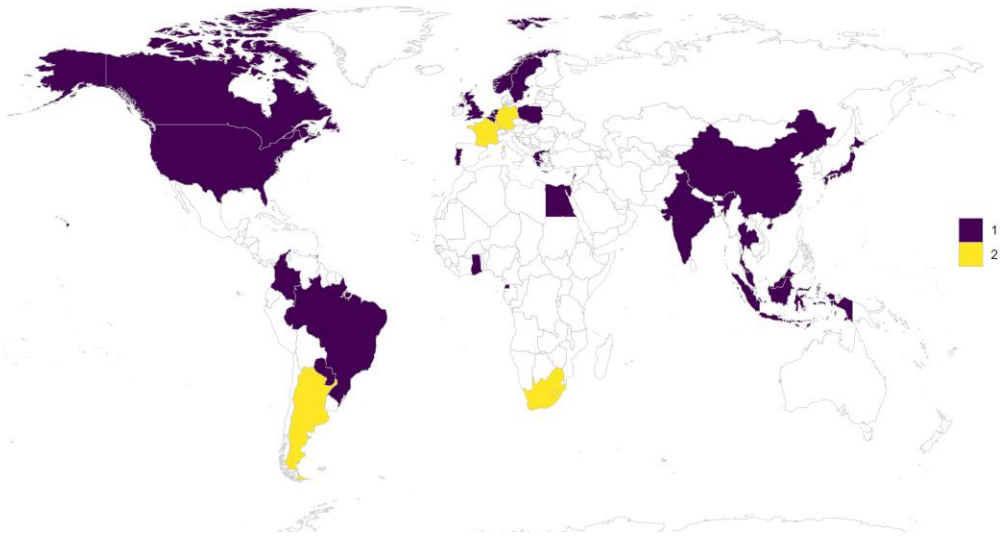


Figure 1. Location of identified metascience observatories ("2" indicates that two observatories were in this country).

Seven candidates were excluded due to a lack of consistent dissemination of products (4a). These included academies of science, ministries of science, national councils, and research labs—all of which met the initial criteria but failed to have consistent publications. One corporation was excluded at (4b) for not providing public access to their reports.

Nine candidates were discarded for technical difficulties (i.e., the websites were inaccessible). These represented a range of countries (Vietnam, Slovenia, Mozambique, Algeria, Nicaragua, Spain, and Austria) and types of institutions (national academies, governmental institutions, and research projects).

Thirty-nine candidates (19%) met all the inclusion criteria. Of these, 31 were explicitly tied to a country; with 28 unique countries represented. The remaining eight were multinational organizations (two with a focus on Europe, one with a focus on Latin America, and the others global (e.g., World Bank and UNESCO)).

Future work

In the next stage of our work, we will code metascience observatories according to several variables, considering the type of organization (e.g., NGO, university, government), the level of autonomy it has in the conduct of business, the type of data collected and produced, the core functions and scope of work, and the intended audiences of the observatory. We are particularly interested in the degree to which these observatories serve in a capacity to inform science diplomacy. Therefore, our next stage of analysis will move from the unobtrusive to obtrusive, meeting with directors and staff of the observatories to understand their functions and context more thoroughly. The goal will be to both describe metascience observatories, but also to be able to provide guidance to current and future observatories on the ways in which they can have heightened relevance both nationally and globally.

Acknowledgments

We are grateful to all members of the IMSO4DIPLO project team, who have contributed in various ways to this work, through the submission of the proposal and discussions in general team meetings. We also acknowledge the IMSO4DIPLO funding from each of our countries: United States (National Science Foundation #2437013); Poland (National Science Centre UMO-2023/05/Y/HS2/00201); South Africa (TAPG231103160452); Brazil (São Paulo Research Foundation. – FAPESP #23/15178-1).

References

AAAS (2010). *New frontiers in science diplomacy—navigating the changing balance of power*. The Royal Society.

Durkheim, E. (1983/1997). *The Division of Labor in Society*. New York: Free Press.

Ezrahi, Y. (2012). *Imagined Democracies: Necessary Political Fictions*. Cambridge University Press.

Flink, T. & Schreiterer, U. (2010). Science diplomacy at the intersection of S&T policies and foreign affairs: towards a typology of national approaches. *Science and Public Policy*, *37*(9), 665–677.

Jasanoff, S. (2009). Essential parallel between science and democracy. *Seed Magazine*.

Macaulay, S. (1966). *Law and the Balance of Power: The Automobile Manufacturers and Their Dealers*. New York: Russel Sage Foundation.

Macedo, D. J., & Maricato, J. de M. (2022). Observatorios de CTI: conceptos, servicios, indicadores y fuentes de información. *Revista Iberoamericana De Ciencia, Tecnología Y Sociedad - CTS*, *17*, 36–60.

Scherer, F. (1964). *The Weapons Acquisition Process: Economic Incentives*. Boston: Division of Research, Graduate School of Business, Harvard University.

Stinchcombe, A. (1990). *Information and Organizations*. University of Chicago Press.

Thorpe, C. (2013). Science and scientific knowledge. In: Runehov, A.L.C., Oviedo, L. (eds) *Encyclopedia of Sciences and Religions*. Springer, Dordrecht.

Weber, M. (1922/1978). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press.

Weisskopf, V. F. (1969). The privilege of being a physicist. *Physics Today*, 22(8): 34.

Williamson, O. (1975). *Markets and Hierarchies*. New York: Free Press.

https://doi.org/10.51408/issi2025_130

Country Self-Preference and National Research Systems: A Path to Independence or Isolation?

Jianjian Gao[1], Alexander J. Gates[2]

[1]psp2nq@virginia.edu, [2]agates@virginia.edu
School of Data Science, University of Virginia, Charlottesville, Virginia (USA)

Abstract

Scientific research is shaped by the interplay between national priorities and international collaboration, essential for advancing global knowledge. Although international collaboration improves research quality and impact, it may reduce national visibility and weaken domestic research networks. Conversely, self-reliant research ecosystems, often reflected in country self-citation rates, can promote research independence and address local challenges but may limit access to global resources. This study examines the balance between these dynamics using bibliometric data from OpenAlex, covering 264 countries from 1960 to 2023. We operationalize country self-preference by analyzing the proportion of citations countries give to their own work and measure international collaboration through the fraction of co-authored publications. The quality of national research ecosystems is assessed by the share of publications in top journals. Fixed-effect panel regression reveals that while international collaboration consistently boosts research quality, national self-preference also positively contributes when balanced effectively with collaboration. Our findings highlight the nuanced strategies nations employ to strengthen their research ecosystems, demonstrating that research independence and global collaboration can be complementary. This work provides actionable insights for policymakers seeking to optimize national scientific performance while fostering equitable and impactful international partnerships.

Introduction

Scientific research operates within a complex interplay of national priorities and international collaboration, both of which are critical for advancing global knowledge and innovation (Marginson, 2022). These dual imperatives often create tensions (Mormina, 2019; Harden-Davies and Snelgrove, 2020); while an interconnected global scientific ecosystem may heighten the overall productivity and efficiency of global science, it potentially diminishes national visibility and domestic research networks (Wagner et al., 2015). Furthermore, some countries, such as Iran, face exclusion from international collaboration due to geopolitical restrictions, forcing them to develop their research infrastructure and scientific capacity with limited external support and partnerships.

Here, we aim to uncover patterns in how nations navigate the trade-offs between fostering self-reliant research ecosystems and engaging in the global scientific enterprise. This question addresses a significant gap in current understanding, as existing literature predominantly focuses on promoting international collaboration without adequately considering its impact on national research ecosystems.

International collaboration has become a cornerstone of modern science, enabling the pooling of resources, expertise, and diverse perspectives to address complex global challenges (Wagner et al., 2001; Adams, 2012). Research consistently underscores

the advantages of cross-border scientific partnerships, as internationally co-authored publications tend to achieve higher citation rates and broader cross-disciplinary impact compared to domestic-only collaborations (Wagner and Jonkers, 2017; Adams, 2013; Glanzel and Schubert¨ , 2001). Over the past three decades, the scale and scope of international collaboration have expanded remarkably (Wagner and Leydesdorff, 2005; Leydesdorff and Wagner, 2008; Chen et al., 2019). Recognizing these benefits, policymakers increasingly prioritize international partnerships in research funding strategies (Katz and Martin, 1997). As Wagner et al. (2015) describe, the global research network is emerging as a new organizational structure that complements—and in some cases supersedes—traditional national systems. For developing countries, international collaboration often serves as a critical mechanism for building national scientific capacity (Harris, 2004). However, despite its many advantages, global collaboration networks remain unequal. Researchers from higher-income countries frequently dominate partnerships, shaping research agendas and benefitting disproportionately (Glanzel and Schubert¨, 2001). Furthermore, geopolitical tensions, funding limitations, and language barriers present significant obstacles to equitable participation in international science, underscoring the need for policies that foster more inclusive and sustainable collaboration frameworks.

Country self-citation offers valuable insights into national research ecosystems, reflecting the extent to which nations rely on and build upon their domestic scholarly contributions (Bakare and Lewison, 2017; Shehatta and Al-Rubaish, 2019; Baccini et al., 2019; Baccini and Petrovich, 2023). While often criticized as a sign of insularity or bias—potentially inflating metrics like the h-index and journal rankings—self-citation can also signify research independence and the ability to address local challenges, particularly in maturing scientific systems (Lariviere et al., 2018; Ladle et al., 2012). Various metrics, such as the self-citation rate and over-citation ratio, attempt to quantify this phenomenon, though recent approaches like fractional citation counts aim to reduce biases related to country size (Qiu et al., 2024). While self-citation often increases alongside international collaboration within countries, its prevalence varies globally, with higher rates in developing nations reflecting localized research priorities or limited visibility, whereas lower rates in developed countries signify greater integration into global networks (Baccini et al., 2019).

In this study, we critically examine these two pivotal dimensions that significantly influence the scientific performance of nations: country self-preference in citations and international collaboration. Specifically, we use bibliometric data from OpenAlex and operationalize country self-preference by analyzing the distribution of citations a country gives to itself relative to all other countries, employing the Area Under the Receiver Operating Characteristic Curve (AUC) and stratified bootstrap to control for a publication's journal. We next measure international collaboration through the fraction of publications involving international co-authors. Lastly, the quality of a nation's scientific ecosystem is measured by the proportion of its articles published in top journals. We then examine the intricate interplay between country self-preference and international collaboration in driving publications in top journals,

using fixed-effect panel regression to uncover their combined impact. Additionally, we explore pathways for strengthening scientific capacity, identifying level sets which reflect the trade-offs between international collaboration or bolstered domestic research infrastructure. Understanding these dynamics is critical for shaping global science policy, as it highlights the nuanced and emergent strategies adopted nations to optimize their research ecosystems and enhance their contributions to the global scientific enterprise. By identifying key drivers and trade-offs, this work provides actionable insights for policymakers to foster equitable and impactful global collaboration while supporting the sustainable development of national research systems.

Data and Methods

We leverage bibliometric data drawn from the OpenAlex bibliometric database in July 2022. We used all indexed "journal-article" and "proceedings-article" records listed as published after 1900 and excluded any publication that did not list an institutional address. Publications are associated with countries using the institutional addresses listed by the authors. We assign a full unit credit of a publication to every country of affiliation on the paper's author byline ("full counting"). In addition, we control for the influence of author self-citation (Aksnes, 2003) and institution self-citation (Wuestman et al., 2019) by removing all citations between publications that share at least one author or at least one affiliation.

We use data on national GDP and percentage of R&D investment from the World Bank to approximate the economic wealth and size of each country. The dataset covers 264 countries from 1960 to 2023.

Fraction with International Authorship

For each country in each year, we count the fraction of publications that share at least one authorship with at least one other country (international collaboration).

National citation preference

We fix a year y, a source country (citing country) s and a target country t (cited country). We then find all publications $n_{s,y}$ worldwide published in year y that also received citations from the source country's 5-year publications. Then, we focus on a target country t, identifying a subset of $n_{s,t,y}$ publications within $n_{s,y}$ publications. We produce 100 stratified bootstrap samples from the worldwide distribution such that the number of publications in each journal exactly matches the counts observed in the target country, thus controlling for both disciplinary differences in citation and one sense of scientific quality.

Finally, we use the Area Under the receiver-operator Curve (AUC) as a measure of the extent to which the cited country's publications ($n_{s,t,y}$) are randomly distributed throughout the citing country's ($n_{s,y}$) ranking. The AUC is a measure of the probability (between 0 and 1) that a randomly chosen publication from the cited country is ranked higher than a randomly chosen publication from any other country; a value of 1 reflects the cited country's publications are over-expressed towards the

top of the ranking, 0 occurs when the cited country's publications are under-expressed towards the bottom of the ranking, and 0.5 denotes a random distribution throughout the ranking. In this study, we set the source country *s* and the target country *t* to be the same country, and obtain the AUC for country self-preference (See Gates et al. (2024) for more details).

We can further quantify the statistical significance of the over/under-representation of a specific country in the citation counts due to the equivalence of the AUC and Mann-Whitney U statistic (DeLong et al., 1988; Sun and Xu, 2014).

Normalized Fraction of Top Journal Articles

To capture the quality of a nation's scientific ecosystem, we use a normalized measure for the fraction of their articles appearing in top journals. Specifically, for each journal in each year, we take the mean log of citations over 5 years to each of its articles. We then rank the journals with publications in each of the 252 subfields in OpenAlex, and take the top 50 journals by subfield; the union set over subfields represents our top journal selection. We then create a normalized measure by dividing by the total number of articles in those journals in that year, thus controlling for the variation in publication volume.

International collaboration and domestic research capacity

We first quantify the relationship between the scientific strength of a nation and the strength of its international collaboration or self-citation preferences. These phenomena, though often studied separately, are intricately linked to the broader dynamics of the scientific capacity of nations.

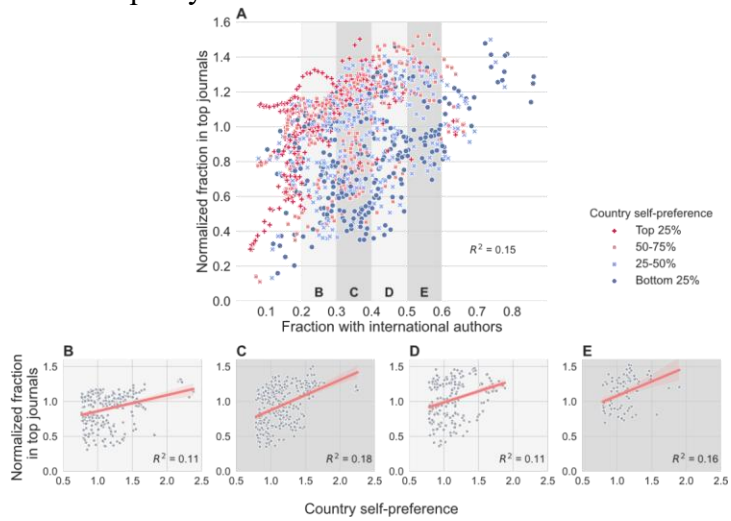


Figure 1: Panel A shows the relationship between the fraction of publications. Panels B-F detail conditional relationships within specific international publication fractions, with corresponding regression lines annotated by R-square (R^2). The shaded area around the regression line represents the 95% confidence interval.

As shown in Fig.1A, there is a strong positive correlation between the fraction of publications with international authors and the normalized fraction of publications in top journals, emphasizing the critical role of international collaboration in enhancing research impact. This corroborates the conclusion by Wagner and Jonkers (2017) that "open countries have strong science". However, the level of openness alone cannot fully explain the variation in high-impact publication rates. Among countries with similar levels of international collaboration, the normalized fraction of top articles varies significantly, suggesting that other factors contribute to research success. The impact of country self-preference, as a proxy of country's scientific independence, on top journal performance is not uniform. When examining countries with similar levels of internationalization (30-40% or 40-50% international authorship), a clearer positive linear trend emerges (Fig. 1B-E), suggesting that the relationship between self-preference and high-impact publications is more readily observable when comparing countries with similar degrees of international collaboration.

To further elucidate this relations, we use a two-way fixed effect panel regression model in which country self-preference and/or international authorship is used to predict the quality of national scientific outputs in the presence of several common covariates. The regression consistently shows a strong, positive, and highly significant effect across all models for international collaboration ($\beta = 0.1661 \sim 0.1824$, p-value < 0.01), underscoring its critical role in enhancing research quality. Country self-preference also shows positive and significant when controlling for other factors ($\gamma = 0.1021$, p-value < 0.05), suggesting that national autonomy in research can complement collaboration when balanced effectively. High self-preference might reflect a country's capacity for independent research, but its translation into impactful publications improves significantly when coupled with robust international partnerships.

Level-sets of scientific capacity

The interplay between international collaboration and research independence represents a fundamental tension in developing national research systems. Some developing countries rely heavily on international collaborations for training and knowledge-sharing (Harris, 2004), while others struggle with national infrastructure building to solve region-specific problems. This section aims to reveal the tension between these two strategies and their impact on the effectiveness of national research systems.

Fig.2 depicts the relationship between country self-preference and international collaboration as resources for producing high-quality research. For varying levels of scientific quality, indicated by the lines (Q1–Q4, from bottom 25% to top 25% according to the ranking of the normalized fraction of publications in top journals), can be interpreted as different patterns of resource utilization, where higher lines (e.g., Q4) represent more efficient or strategic combinations of these resources, yielding greater normalized fractions of publications in top journals.

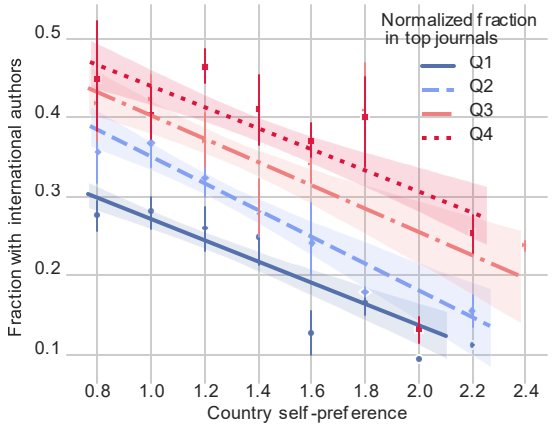


Figure 2: Balance between country self-preference (log of the z-score of the AUC) and international authorship. Shaded areas indicate 95% confidence intervals.

For a fixed quantile of scientific quality, there's a negative trend between country self-preference and the fraction of publications with international authors. This indicates that countries relying on domestic research capacity tend to have fewer external partnerships. This trend reveals the tension between the inclination towards building domestic infrastructure and promotion of international collaboration for national scientific capacity building.

More importantly, this figure demonstrates paths through which countries transition from a lower quality level to a higher one. Countries at lower levels (e.g., Q1) may over-rely on one resource without optimizing the balance due to economic constrains or geopolitical tensions, whereas those on higher levels demonstrate more efficient or strategic resource allocation. To ascend to higher levels, countries must enhance the weaker resources — whether by increasing international collaboration or strengthening domestic research capacity. This analysis underscores the importance of strategic resource utilization and equitable access to collaboration opportunities for achieving research impact. Fostering international partnerships should be complemented by policies to strengthen domestic research infrastructure, support independent researchers, and promote local innovation ecosystems.

Discussion

Our study reveals the complex interplay between international collaboration and country self-preference in scientific research, offering critical insights into national research ecosystem dynamics. Key observations highlight that both international cooperation and self-referential publication practices contribute positively to high-quality research output. The findings have profound implications for science policy. Nations can pursue diverse scientific capacity-building strategies: some may prioritize extensive international networks, while others may focus on strengthening domestic research infrastructures. The observed trends and trade-offs are relevant irrespective of their underlying cause—whether they arise as an emergent properties of the national scientific ecosystem or result from deliberate strategic policy

decisions by governments—highlighting their importance for understanding global research infrastructure. The limitations of this study include potential biases in the OpenAlex dataset and the complexity of measuring the quality of research through the fraction of publications in top journals. Future research could explore individual countries' trajectories of science capacity-building and policies driving the current landscapes in countries.

References

Adams, J. (2012). The rise of research networks. *Nature 490*(7420), 335–336.

Adams, J. (2013). The fourth age of research. *Nature 497*(7451), 557–560.

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics 56*(2), 235–246.

Baccini, A., G. De Nicolao, and E. Petrovich (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS One 14*(9), e0221212.

Baccini, A. and E. Petrovich (2023). A global exploratory comparison of country self-citations 1996-2019. *Plos one 18*(12), e0294669.

Bakare, V. and G. Lewison (2017). Country over-citation ratios. *Scientometrics 113*(2), 1199–1207.

Chen, K., Y. Zhang, and X. Fu (2019). International research collaboration: An emerging domain of innovation studies? *Research Policy 48*(1), 149–168.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988, Sep). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics 44*(3), 837.

Gates, A. J., I. Mane, and J. Gao (2024). The increasing fragmentation of global science limits the diffusion of ideas. Available at arXiv: https://doi.org/10.48550/arXiv.2404.05861

Glanzel, W. and A. Schubert (2001). Double effort= double impact? A critical view at international co-authorship in chemistry." *Scientometrics 50*(2), 199–214.

Harden-Davies, H. and P. Snelgrove (2020). Science collaboration for capacity building: Advancing technology transfer through a treaty for biodiversity beyond national jurisdiction. *Frontiers in Marine Science 7*, 40.

Harris, E. (2004). Building scientific capacity in developing countries. *EMBO reports 5*(1), 7–11.

Katz, J. S. and B. R. Martin (1997). What is research collaboration? *Research policy 26*(1), 1–18.

Ladle, R. J., P. A. Todd, and A. C. Malhado (2012). Assessing insularity in global science. *Scientometrics 93*(3), 745–750.

Lariviere, V., K. Gong, and C. R. Sugimoto (2018). Citations strength begins at home. *Nature 564*(7735), S70– S70.

Leydesdorff, L. and C. S. Wagner (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics 2*(4), 317–325.

Marginson, S. (2022). "All things are in flux": China in global science. *Higher Education 83*(4), 881–910.

Mormina, M. (2019). Science, technology and innovation as social goods for development: Rethinking research capacity building from Sen's capabilities approach. *Science and Engineering Ethics 25*(3), 671–692.

Qiu, S., C. Steinwender, and P. Azoulay (2024, May). Paper tiger? Chinese science and home bias in citations. Working Paper w32468, National Bureau of Economic Research. Available at SSRN: https://ssrn.com/abstract=4833948.

Shehatta, I. and A. M. Al-Rubaish (2019). Impact of country self-citations on bibliometric indicators and ranking of most productive countries. *Scientometrics 120*(2), 775–791.

Sun, X. and W. Xu (2014, Nov). Fast implementation of Delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters 21*(11), 1389–1393.

Wagner, C. S., I. Brahmakulam, B. Jackson, A. Wong, and T. Yoda (2001). Science and technology collaboration: Building capacity in developing countries. *Document Number MR-1357.0-WB, RAND Corporation, Santa Monica, CA*.

Wagner, C. S. and K. Jonkers (2017, October). Open countries have strong science. *Nature 550*(7674), 32–33.

Wagner, C. S. and L. Leydesdorff (2005). Mapping the network of global science: Comparing international co-authorships from 1990 to 2000. *International journal of Technology and Globalisation 1*(2), 185–208.

Wagner, C. S., H. W. Park, and L. Leydesdorff (2015, July). The continuing growth of global cooperation networks in research: A conundrum for national governments. *PLoS One 10*(7), e0131816.

Wuestman, M. L., J. Hoekman, and K. Frenken (2019). The geography of scientific citations. *Research Policy 48*(7), 1771–1780.

https://doi.org/10.51408/issi2025_098

Current Interdisciplinarity Measures Fail to Reflect Authors' Perspectives

Dag W. Aksnes[1], Henrik Karlstrøm[2], Fredrik N. Piro[3]

[1] dag.w.aksnes@nifu.no, [2] henrik.karlstrom@nifu.no, [3] fredrik.prio@nifu.no
NIFU – Nordic Institute for studies in Innovation, Research and Education, Økernveien 9, 0653 Oslo (Norway)

Abstract

The objective of this paper is to compare bibliometrically constructed indexes of interdisciplinarity with authors' self-assessments of the interdisciplinarity of their own papers, thereby providing knowledge into how well these indicators correspond with researchers' own perceptions. The bibliometric interdisciplinarity measures analyzed include the Shannon entropy, ${}^{2}D^{S}$, and DIV* indicators. The data analyzed in the study are derived from two separate questionnaire surveys, in which authors were asked about specific articles they had published. The results reveal that there is little agreement between the bibliometric measures and authors' assessments of interdisciplinarity, with correlations ranging from weak to very weak.

Introduction

Interdisciplinarity or interdisciplinary research (IDR) has become buzzwords in research policy (Cantone, 2024). This is not surprising as interdisciplinarity is increasingly characterizing contemporary research practices (Porter & Rafols, 2009). Moreover, funding agencies are frequently emphasizing IDR, for example through establishing interdisciplinary research centers and programs (Avila-Robinson, Mejia & Sengoku, 2021; Chen et al., 2021).

However, defining exactly what interdisciplinarity means is challenging (Miller, 2020). Several definitions and concepts of IDR exist in the literature (Laursen, Motzer & Andersson, 2022; von Wehrden et al., 2019). At the same time, numerous indicators attempting to measure interdisciplinarity bibliometrically have been developed over the years. While these indicators ostensibly aim to capture the same phenomenon, they often produce divergent results (Avila-Robinson, Mejia & Sengoku, 2021; Cantone, 2024). This means that the degree of interdisciplinarity for any given unit of analysis can vary significantly depending on the specific measures applied. Wang & Schneider fundamentally question the validity of the measures and argue that "the current measurements of interdisciplinarity should be interpreted with much caution" (2019, p. 239).

Against this background there is a need for more validity studies. While all proposed IDR indicators strive for accurate and valid measurement of interdisciplinarity (given the chosen definition), there is a risk that these indicators capture *bibliometric-derived interdisciplinarity* – patterns and connections discernible in bibliometric data – rather than *"real-life"* interdisciplinarity. We believe that it is important to understand how IDR measures align with perceptions of interdisciplinarity from the very producers of the research. Thus, the objective of this paper is to compare bibliometrically constructed indexes of IDR with self-assessments of

interdisciplinarity by the authors of the same papers, thereby providing knowledge into how well these indicators correspond with researchers' own perceptions.

Few prior studies have addressed the issue of construct validity. Zhang et al. (2018) examined over 150,000 PLoS One articles, comparing an IDR measure based on cited references with authors' departmental affiliations, and found a low correlation between the two. Roessner et al. (2013) conducted an ethnographic study of a single researcher, examining how bibliometric measures aligned with perceptions of knowledge integration but did not find a simple answer to the question of the validity of IDR indicators. Of particular relevance to the present study is the work of Avila-Robinson, Mejia, and Sengoku (2021), which, to our knowledge, is the only study that analyses a larger dataset to compare bibliometric IDR measures with authors' self-assessments of interdisciplinarity. Their findings based on analyses of a thousand publications reveal relative weak yet statistically significant associations (coefficients ranging from 0.20 to 0.27) between self-assessments of interdisciplinarity and four IDR measures.

In our study we analyse three indicators which are commonly used to analyse interdisciplinarity:

- Shannon's entropy. Originally introduced as a measure of "information uncertainty", Shannon entropy is often used to quantify the diversity of disciplines referenced in a given publication or set of publications (see e.g Leydesdorff, Wagner, & Bornmann 2019a).
- The ${}^{2}D^{S}$-measure (Zhang, Rousseau & Glänzel (2016) which is related to, but a further development of the Rao-Stirling measure. It quantifies the diversity (the range of disciplines) and disparity of disciplines (their unrelatedness) cited in a research publication.
- DIV*. Leydesdorff, Wagner & Bornmann (2019a) introduced the DIV measure, an alternative measure of interdisciplinarity that combines balance, variety, and disparity into a single metric. This measure was later refined into DIV* to correspond with Rousseau's (2019) principles for interdisciplinarity metrics (Leydesdorff, Wagner & Bornmann, 2019b).

Data & methods

Survey data

The data analyzed in this study are derived from two separate questionnaire surveys, in which authors were asked about specific articles they had published. The surveys addressed various dimensions of the articles. Survey 1 focused on quality aspects and the research process, with interdisciplinarity being one of the dimensions examined (see Aksnes et al. (2023) for survey findings related to research quality and citation rates). Survey 2 focused on publication practices in environmental sciences, where the authors were asked to evaluate the articles and their experiences with the publication process. The survey particularly targeted issues related to interdisciplinarity. Both surveys included numerous questions, but for the purpose of this paper only a few were used.

In Survey 1, authors were simply asked to rate the extent to which each of three specific articles was interdisciplinary, using a scale from 1 (low) to 5 (high). No definition of interdisciplinary was provided, allowing respondents to apply their own interpretations. In Survey 2, the authors were asked to indicate whether the following characterizes the paper: a) Based on research from multiple academic disciplines (multidisciplinary), b) Combines and integrates research across academic disciplines (interdisciplinary). The following alternatives were given: Yes, Partly, No, Not Applicable/Don't Know. In another question they were asked to describe who was involved in the research that the paper builds on. One of the options provided was: "Researchers from other fields were involved in the research". Here a binary response option was used (Selected/Not Selected). Thus, this survey applies short definitions and distinguishes between multidisciplinary and interdisciplinary research.

Survey 1 one was sent to a stratified sample of researchers in Norway. The survey was conducted in January 2022 and questionnaires were distributed to a sample of 1,250 researchers. The survey asked researchers to assess three of their own papers, selected through stratified random sampling based on citation metrics (one paper from each of the following citation rank percentiles: top 10%, 10-50% and 50-100%). With a response rate of 47%, the final sample included 592 researchers, each contributing three publications. As a result, the study encompasses assessments of 1,780 publications, of which 1,695 included responses to the question on interdisciplinarity.

Survey 2 was a global survey conducted in the spring of 2023. It was distributed to corresponding authors of articles in 21 environmental science journals, including the mega-journal *Sustainability*. Out of an initial sample of approximately 14,500 selected authors, the survey achieved a response rate of 12.5%, resulting in 1,800 responses. Consequently, the survey encompasses the evaluation of 1,800 articles, of which approximately 1,510 included responses to the questions on interdisciplinarity.

Bibliometric data

The study relies on bibliometric data from the Web of Science (WoS) database, which has been used to calculate bibliometric interdisciplinarity measures. We applied a local version of WoS maintained by the Norwegian Agency for Shared Services in Education and Research. We applied data on the references of the publications to calculate the three interdisciplinarity scores used in the study. Only publications with at least 10 WoS-indexed references were included in the comparative analysis, as a minimum number of references is required to reliably calculate interdisciplinarity scores. This threshold reduced the number of articles by approximately 7%, from 1,559 articles in Survey 1 to 1,756-1,455 in Survey 2 (depending on question).

Results

As previously described, we have two different surveys addressing the interdisciplinarity of publications. Below we report the results of both of them. The analysis reveals a limited correspondence between the authors' assessments of

interdisciplinarity and the bibliometric interdisciplinarity scores. Figure 1 shows a Box-Whisker plot comparing author ratings of interdisciplinarity (survey 1) with DIV*-values. As can be seen there is a tendency that publications rated as having high interdisciplinarity generally have higher DIV* values. For example, publications rated with the lowest level of interdisciplinarity (1) have a median DIV* value of 0.021, whereas those rated at the highest level (5) obtained a median DIV* value of 0.027. However, the most striking result is the large disparity in results. Many publications rated by authors as highly interdisciplinary do not exhibit high scores on the bibliometric measures, and vice versa. Very similar plots and patterns were observed in Survey 2, and to avoid redundancy, separate figures are not presented.

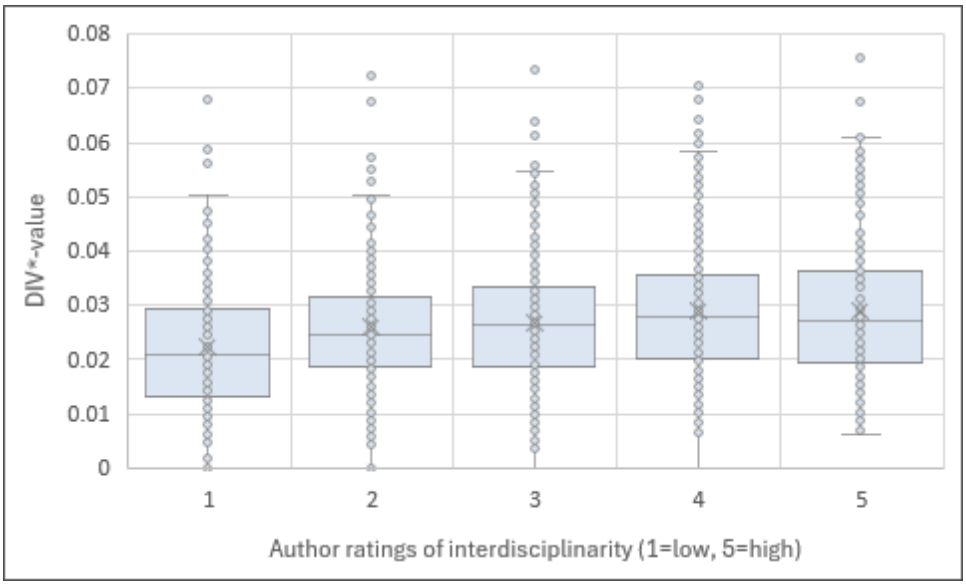


Figure 1. Box-Whisker plot of author ratings of interdisciplinarity (survey 1) and DIV*-values (interquartile range (1st-3rd), mean (cross), median (line within the box).

Table 1 shows the results of a correlation analysis using Spearman's rank-order method, meaning it is based on ranks rather than raw values. This is due to survey results measured on an ordinal scale, specifically for Survey 2, the variables are either trinary (e.g., Yes/Partly/No) or binary (e.g., Selected/Not Selected), necessitating the use of a non-parametric method.

The results on Survey 1 and the results on two of the questions in Survey 2 (Research from multiple academic disciplines (multidisciplinary) - Combines and integrates research across academic disciplines (interdisciplinary) are very similar. For these variable Spearmans's rho is in the range of 0.130-0.175 for all three bibliometric measures. This is a weak/very weak correlation. However, the p-values for these correlations are extremely low ($p < 0.00001$), indicating strong statistical significance, partly explained by the large number of observations. Among the associations, the correlations with Integrated disciplines are slightly higher, suggesting that this dimension aligns more strongly with the interdisciplinarity

measures. Conversely, the coefficients for the DIV* metric are marginally higher than those for the other measures, suggesting that DIV* is slightly more reflective of the authors' views.

Table 1. Correlation analysis: Relationship between interdisciplinarity measures and the authors' assessments (Spearman's rank correlation coefficients and p-values).

	Shannon entropy		$^2D^S$		*DIV**		
	Spear-man's rho	*P-value*	*Spear-man's rho*	*P-value*	*Spear-man's rho*	*P-value*	*N*
Interdisc. rank (1)	0.152	1.5×10^{-9}	0.130	2.6×10^{-7}	0.160	2.1×10^{-14}	1559
Multiple disc. (2)	0.160	7.3×10^{-10}	0.155	2.2×10^{-9}	0.158	1.1×10^{-9}	1466
Integrated disc. (2)	0.173	3.0×10^{-11}	0.155	2.8×10^{-9}	0.175	1.7×10^{-11}	1455
Other field researchers (2)	0.056	0.019	0.027	0.25	0.053	0.027	1756

For the indicator "Other field researchers," correlations are generally negligible (rho < 0.1) and have higher p-values. While the correlation with Shannon entropy and DIV* is marginally significant ($p < 0.05$), the correlation with $^2D^S$ is not statistically significant ($p = 0.25$). This suggests that the interdisciplinarity measures are less relevant for capturing this dimension.

Discussion & conclusions

The findings of this study contribute to the ongoing debate regarding the validity of bibliometric measures of interdisciplinarity by juxtaposing these with researchers' self-assessments. The results reveal that there is little agreement between bibliometric measures (Shannon entropy, $^2D^S$, and DIV*) and authors' assessments of interdisciplinarity: the correlations are weak to very weak. This suggests that bibliometric measures only to a very little extent capture researchers' perceptions of the interdisciplinarity of their own work. In particular, we got poor correspondence for the dimension "Other field researchers," indicating that bibliometric measures are not capable of reflecting the involvement of researchers from different fields. These findings are consistent with earlier studies, such as those by Avila-Robinson, Mejia, and Sengoku (2021), which also reported relatively weak correlations between bibliometric interdisciplinarity metrics and authors' views, albeit slightly stronger than observed in our study.

Another notable finding is that, despite variations in their construction and methodological foundation, the three bibliometric measures examined show a similar level of correspondence. The DIV* metric appears to align marginally better with

authors' self-assessments compared to Shannon entropy and ${}^{2}D^{S}$. This is reflected in the slightly higher correlation coefficients for DIV* across most variables.

The study is framed as a kind of validation exercise of bibliometric interdisciplinarity measures.

Our findings suggest that these metrics have inherent limitations and lack validity, at least insofar as researchers' perceptions can be used as benchmarks to assess the issue. This divergence naturally leads to the conclusion that one should not rely solely on bibliometric indicators to assess interdisciplinarity.

Nevertheless, the issue is complex. The relationship between bibliometric metrics and author assessments can also be interpreted in the opposite direction. Systematic and objective measures of interdisciplinarity are compared with subjective assessments by the authors. After all, measures like Shannon entropy, 2DS, and DIV* might provide useful insights into the diversity and integration of disciplines referenced in a publication. However, this dimension does not seem to reflect the "real-life" interdisciplinarity experienced by researchers. Perhaps researchers emphasize a more composite set of factors in their assessments, such as disciplinary norms, collaboration dynamics, methodological and theoretical approaches, etc. This suggests that researchers' perceptions are likely influenced by factors not readily captured by the bibliometric indicators, leading to limited comparability. Generally, the two approaches can be expected to correlate positively only if the aspects assessed by the authors correspond to those reflected in the bibliometric metrics.

Moreover, authors' assessments of their own publications may not be entirely inter-subjective. Different authors could rank the interdisciplinarity of the same publication differently due to varying perspectives, biases, or even memory limitations.

Thus, more studies are needed to determine the extent of inter-subjectivity in authors' assessments of interdisciplinarity, as well as to explore factors that shape researchers' perceptions of interdisciplinarity. Future studies could also expand the range of bibliometric measures to assess whether alternative metrics align better with researchers' perceptions (see, e.g. Abramo, D'Angelo & Di Costa (2012).

Acknowledgments

This work was supported by the Research Council Norway (RCN) [grant number 256223] (R-QUEST).

References

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. *Journal of the American Society for Information Science and Technology*. 63: 2206-2222.

Aksnes, D. W., Piro, F. N., & Fossum, L. W. (2023). Citation metrics covary with researchers' assessments of the quality of their works. *Quantitative Science Studies*, 4(1), 105-126.

Avila-Robinson, A., Mejia, C. & Sengoku, S. (2021). Are bibliometric measures consistent with scientists' perceptions? The case of interdisciplinary research. Scientometrics, 126, 7477-7502.

Cantone, G.G. (2024). How to measure interdisciplinary research? A systemic design for the model of measurement. *Scientometrics*, 129, 4937-4982.
Chen, S., Qiu, J., Arsenault, C. & Lariviere, V. (2021). Exploring the interdisciplinarity patterns of highly cited papers. *Journal of Informetrics*, 15, 101124.
Chen, S., Zhang, K., Qiu, J. & Chai, J. (2024). Interdisciplinary and expert rating: an analysis based on faculty opinions. *Scientometrics*, 129, 6597-6628.
Laursen, B.K., Motzer, N. & Anderson, K.J. (2022). Pathways for assessing interdisciplinarity: A systematic review. *Research Evaluation*, 31(3), 326-343.
Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019a). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal Informetrics*, 13 (1), 255–269.
Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019b). Diversity measurement: Steps towards the measurement of interdisciplinarity? *Journal Informetrics*, 13(3), 904–905.
Miller, R. (2020). Interdisciplinarity: Its Meaning and Consequences. *Oxford Research Encyclopedia of International Studies*. Retrieved 10 Oct. 2024, from https://oxfordre.com/internationalstudies/view/10.1093/acrefore/9780190846626.001.0001/acrefore-9780190846626-e-92.
Porter A. L. & Rafols I. (2009). Is Science Becoming More Interdisciplinary? Measuring and Mapping Six Research Fields over Time'. *Scientometrics*, 81, 719–45.
Roessner, D., Porter, A.I., Neresessian, N.J. & Carley, S. (2013). Validating Indicators of Interdisciplinarity: Linking Bibliometrics Measures to Studies of Engineering Research Labs. *Scientometrics*, 94, 439-468.
Rousseau, R. (2019). On the Leydesdorff-Wagner-Bornmann proposal for diversity measurement. *Journal Informetrics*, 13(3), 906.
Wang, Q. & Schneider, J.W. (2019). Consistency and validity of interdisciplinary measures. *Quantitative Science Studies*, 1, 239-263.
von Wehrden, H., Guimaraes, M.H., Bina, O., Varanda, M., Lang, D.J., John, B., Gralla, F., Alexander, D., Raines, D., White, A. & Lawrence, R.J. (2019). Interdisciplinary and transdisciplinary research: finding the common ground of multi-faceted concepts. *Sustainability Science*, 14(6).
Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal Association for Information Science Technology*, 67, 1257–1265
Zhang, L., Sun, B., Chinchilla-Rodriguez, Z. & Huang, Y. (2018). Interdisciplinarity and collaboration: On the relationship between disciplinary diversity in departmental affiliations and reference lists. *Scientometrics*, 117, 271-291.

https://doi.org/10.51408/issi2025_108

Difference between Preprint and Journal Systems

Chiaki Miura[1], Ichiro Sakata[2]

[1]t.miura@gnt.place,[2]isakata@ipr-ctr.t.u-tokyo.ac.jp
Faculty of Engineering, The University of Tokyo (Japan)

Abstract

Preprints are considered to supplement journal-based systems for the rapid dissemination of relevant scientific knowledge. Emerging frame works such as the publish-review-curate (PRC) model, post-publication peer review, and diamond open access collectively signal a shift towards preprint-led academic norms. The preprint system has historically been supported by evidence showing no significant differences in semantics, teaming, referencing, or quality control between preprints and published reports. However, as preprints increasingly serve as independent mediums for scholarly communication rather than precursors to the version of record, it remains uncertain how these emerging norms will impact wider scholarly practice.

This paper provides insights into how these norms might evolve by analyzing the differences between preprints and journal articles, highlighting their implications for the future of scholarly communication. We examined the use, contributors, and epistemic networks of preprints. Surprisingly, preprint citations have a larger imbalance, indicating the effect that actors disproportionately rely on reputable peers in an unvetted environment. Contributor shares for preprints are consistent between preprint-only and preprints with subsequent publication, differing from journal trends. Research institutes and non-profits have a higher share of preprints, while companies stand out as an exception, with a notable tendency to focus on preprint-only papers. Future research will benefit from natural experiments that enable direct comparisons and more detailed data on academic practices within preprint systems.

Introduction

The increasingly rapid transformations in modern society, coupled with the growing role of science, have elevated the importance of the rapid dissemination of scientific findings. Preprint is intended to minimize the publishing delay due to article processing (Goldschmidt-Clermont, 2002) and has garnered significant attention during the COVID-19 pandemic, stimulated considerable debate if preprints can be cited and relied upon as concrete evidence for life (Kwon, 2020).

Although the major concern with preprints has been that only a fraction of them are qualified and thus considered not to undergo the established scrutiny (Sheldon, 2018), when viewed from content, there is growing evidence that preprints can match journal articles. Compared to the corresponding version of the record, preprints show no significant difference in reference (Akbaritabar et al., 2022), authorship (Brierley et al., 2022), and qualitative expert evaluation (Carneiro et al., 2020). A considerable proportion of preprints undergo peer review, with about two-thirds of whole preprint submissions in every publish-year cohort eventually published in journals (Table.1 cf. Fraser et al. (2020)); major publishers have begun officially including preprints in citation indices (Elsevier, 2021), making preprints role less distinguishable with journals'.

However, there remains a significant gap in understanding how the rise of preprints may transform scholarly practices. Prior research focused on descriptive

characteristics of preprint as a precursor in relation to journal articles. The emergent peer-review models like post-publication peer review platforms such as eLife and F1000, and publish-review-curate model (Eisen et al., 2020) such as metaRoR consider preprint as an independent, main medium of academic discourse, along with the rise of the peer review pipeline that processes and verifies articles on preprint servers (Weissgerber et al., 2021).

Table 1. Top ten major journals bioRxiv preprints are subsequently published between 2013 and 2024. eLife articles are excluded from journal.

Journal	*Publication*
Nature Communications	6,074
PLOS ONE	5,501
Scientific Reports	4,535
Proceedings of the National Academy of Sciences	3,100
PLOS Computational Biology	2,448
Bioinformatics	1,993
Cell Reports	1,816
Nucleic Acids Research	1,683
NeuroImage	1,427
PLOS Genetics	1,378
Top 10 total	29,955
All bioRxiv Preprints	268,470

This study aims to address this gap by examining differences in academic practices between preprints and journal-based systems, comparing the two systems from three perspectives: the use, contributors, and epistemic network.

Especially we focus on the imbalance and bias in citation practices of researchers. It is known that in an environment where actors do not have prior knowledge about the validity of information, they disproportionately rely on reputable peers (Bendtsen et al., 2013). This can promote imbalance and hinder new theories and practices from taking over, impeding science progress (Chu and Evans, 2021). Reference lists in one article are often directly transported to another (MacRoberts and MacRoberts, 1989), which may leave traces in citation distribution differently from other propagation of reference preference. Cultural diffusion model explains conforming frequency-dependent copying significantly deforms the power-law distribution of

traits frequency (Mesoudi and Lycett, 2009). Citation network citing to and within the preprint system, mapped to the journal system via semantic similarity, can reveal the hidden selection bias in the system.

Method and Materials

In the following section, we use the term curate to refer to the act of making preprints available in a journal, article and to publish to refer to any of preprints or journal articles indifferently, and the act of making them available, respectively. We selected biology and the medical field as our analysis of interest, although it is notable that later we further confirm the robustness with other fields with independent datasets. We combined the world's largest bibliographic database, OpenAlex, with the snapshot of the largest preprint server, bioRxiv, supplemented and validated by journal publication data from Scopus. We collect 268,470 OpenAlex records of preprint articles published from Jan. 2013 to Dec. 2024, which matches 137,011 curated preprints and 131,459 non-curated preprints on bioRxiv.

Journal ages are inferred from the first year with a noticeable publication threshold N, where we took N = 30 for our analysis. Topic coverage is calculated based on variety, namely the unique number of Scopus ASJC topic categories assigned to at least M articles, where we simply considered the case M = N. All the analyses below consider journal articles published between 2015 and 2020 unless stated otherwise. This is to eliminate the effect of citation inflation and other year fixed effect, as well as the effect of COVID-19-related preprints. In the same way, citation is the count five years after publication.

Result

Longitudinal citation count of an article grows exponentially due to the preferential attachment (Jeong et al., 2003). Thus, mere skew does not indicate the presence of reputation bias. Therefore, we first examined the baseline imbalance in journal system.

We measured imbalance by the Gini coefficient, which is suitable for the purpose as it is size agnostic, robust to extreme outliers, and normalized, al though the metric should be interpreted carefully as the same value can result from different curves. Notably, citation distribution within journal is typically lognormal in both journals and preprints (Wang et al., 2013; Fraser et al., 2020). We took the logarithm of the publication volume and citation to address the issue of high variability within the two variables. This transformation helps to normalize the distribution, reduce the impact of extreme values, and make relationships more clearer. Table.2 shows pairwise Pearson correlation between variables. It is important to interpret these correlations with caution as they do not account for any confounding variables.

Table 2. Descriptive statistics and Pearson correlations between variables. Asterisk(*) indicates that the variable is transformed by base ten logarithms.

	mean	*(*	*S.D.*	*)*	*min*	-	*max*	*1*	*2*	*3*	*4*
1. Count	3.4	(	0.3	)	3.00	-	5.52				
2. Average Citation	0.9	(	0.4	)	-0.54	-	2.27	.242			
3. Diversity	2.3	(	1.3	)	0	-	13	.050	.215		
4. Journal Age τ	32.8	(	11.2	)	7	-	54	.300	-.130	-.041	
5. Imbalance G	0.6	(	0.1	)	0.38	-	0.90	-.138	-.398	-.129	.304

Controlling confounding variables, journal age and imbalance significantly positively correlate ($R = 0.313$, $p < 0.001$). This means that even if compared within the same cohort of articles published in the same period, older journals have a higher article presence inequality at the same citation age, indicating that established journals tend to associate with certain canonical groups of works.

In fig. 1, we plotted preprint data on the journal baseline. BioRxiv, with an age $\tau = 13$ years, shows a significant citation imbalance ($G = 0.683$) for curated preprints. Similarly, bioRxiv preprints that remain un-curated within the observed period show a comparable imbalance ($G = 0.710$). This result is surprising, as curated preprints typically have a "cut-off" date after which citations should predominantly accrue to the journal version of the article. Moreover, the majority of biology preprints undergo processing and become available as journal articles within a year (Xie et al., 2021).

This raises the question of whether the observed imbalance is driven by reputable authors disproportionately attracting citations or by other systemic factors. We compare the authors' reputations in journals with their relative impact in preprints. This is a research-in-progress paper, and further research should be done in the near future.

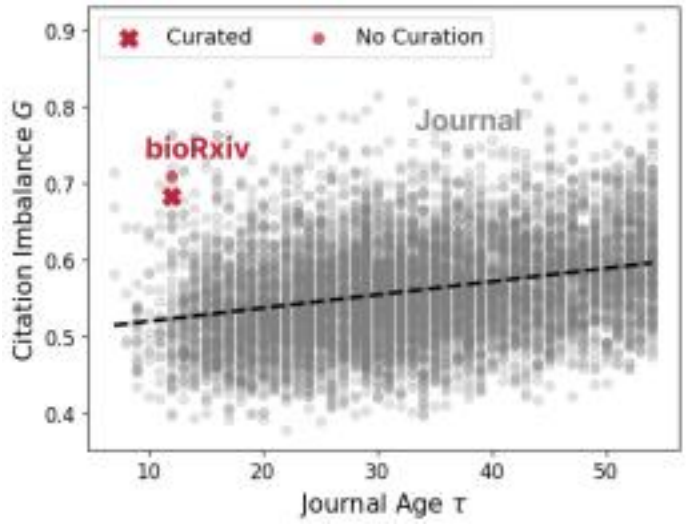


Figure 1. Correlation between Journal age and five-year citation imbalance within each journal. Each point represent one journal. Newer journals tend to have a lower Gini coefficient. Controlling citation inflation does not affect the result.

This raises the question of whether the observed imbalance is driven by reputable authors disproportionately attracting citations or by other systemic factors. We compare the authors' reputations in journals with their relative impact in preprints. This is a research-in-progress paper, and further research should be done in the near future.

Discussion

Our preliminary result shows preprints in biology exhibit significantly higher skew within the source compared to their journal counterparts. This imbalance is not necessarily the result of systemic reputation bias; it may come from other factors, such as the sources accepting risky and potentially innovative ideas and attracting higher quality than average publishing sources. Similar trends in other distinguished journals highlightsthe need for more refined metrics to assess the imbalance and close-up understanding of what contributes the imbalance.

Furthermore, in-depth analysis of scholarly communication in the fields where preprints are already dominant, such as computer science, can enhance the understanding of the new norm.

As initiatives like the PRC model gain traction, scholarly communication is expected to shift from a static publication system to a dynamic process of discourse building, supported by a preprint-centered academic infrastructure. In such a system, scholarly outputs are continuously revised, debated, and reassessed. Maintaining the reliability of this evolving framework requires mechanisms that account for retractions and corrections. For instance, if a preprint is retracted, a corresponding alert should be propagated to all citing papers to prevent the continued dissemination of unreliable findings.

References

Akbaritabar, A., Stephen, D., & Squazzoni, F. (2022). A study of referencing changes in preprint-publication pairs across multiple fields. Journal of Informetrics, 16(2), 101258.

Bendtsen, K. M., Uekermann, F., & Haerter, J. O. (2013). The expert game—Cooperation in social communication (No. arXiv:1312.6715). arXiv.

Brierley, L., Nanni, F., Polka, J. K., Dey, G., Pálfy, M., Fraser, N., & Coates, J. A. (2022). Tracking changes between preprint posting and journal publication during a pandemic. PLOS Biology, 20(2), e3001285.

Carneiro, C. F. D., Queiroz, V. G. S., Moulin, T. C., Carvalho, C. A. M., Haas, C. B., Rayêe, D., Henshall, D. E., De-Souza, E. A., Amorim, F. E., Boos, F. Z., Guercio, G. D., Costa, I. R., Hajdu, K. L., van Egmond, L., Modrák, M., Tan, P. B., Abdill, R. J., Burgess, S. J., Guerra, S. F. S., … Amaral, O. B. (2020). Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. Research Integrity and Peer Review, 5(1), 16.

Chu, J. S. G., & Evans, J. A. (2021). Slowed canonical progress in large fields of science. Proceedings of the National Academy of Sciences, 118(41), e2021636118.

Eisen, M. B., Akhmanova, A., Behrens, T. E., Harper, D. M., Weigel, D., & Zaidi, M. (2020). Implementing a "publish, then review" model of publishing. eLife, 9, e64910.

Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. Quantitative Science Studies, 1(2), 618–638.

Goldschmidt-Clermont, L. (2002, March). Communication Patterns in High-Energy Physics. High Energy Physics Libraries Webzine, 6.

Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. Europhysics Letters, 61(4), 567.

Kwon, D. (2020). How swamped preprint servers are blocking bad coronavirus research. Nature, 581(7807), 130–131.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40(5), 342–349.

Mesoudi, A., & Lycett, S. J. (2009). Random copying, frequency-dependent copying and culture change. Evolution and Human Behavior, 30(1), 41–48.

Preprints are now in Scopus! | Elsevier Scopus Blog. (2021, January 28).

Sheldon, T. (2018). Preprints could promote confusion and distortion. Nature, 559(7715), 445–445.

Turner, S. (2024). bioRxiv preprint and publication details, 2014-2023 [Dataset]. Zenodo.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. Science,

Weissgerber, T., Riedel, N., Kilicoglu, H., Labbé, C., Eckmann, P., ter Riet, G., Byrne, J., Cabanac, G., Capes-Davis, A., Favier, B., Saladi, S., Grabitz, P., Bannach-Brown, A., Schulz, R., McCann, S., Bernard, R., & Bandrowski, A. (2021). Automated screening of COVID-19 preprints: Can we help authors to improve transparency and reproducibility? Nature Medicine, 27(1), 6–7.

Xie, B., Shen, Z., & Wang, K. (2021). Is preprint the future of science? A thirty year journey of online preprint services (No. arXiv:2102.09066). arXiv.

https://doi.org/10.51408/issi2025_120

Disciplinary Identity in the Origins of the Science of Science

Emanuel Kulczycki[1], Przemysław Korytkowski[2]

[1]*emanuel@ekulczycki.pl*
Adam Mickiewicz University in Poznań, Scholarly Communication Research Group, Poznań (Poland)
West Pomeranian University of Technology in Szczecin, Szczecin (Poland)

[2]*pkorytkowski@zut.edu.pl*
West Pomeranian University of Technology in Szczecin, Szczecin (Poland)
National Information Processing Institute, Warsaw (Poland)

Abstract

This paper explores the disciplinary identity of the Science of Science (SoS) in Poland from its inception in 1918 to 2020. The study analyzes over 9,000 articles from three key Polish SoS journals to assess whether the thematic areas proposed by Maria and Stanisław Ossowski in the 1930s remain relevant for categorizing the field. Our findings indicate that while practical-organizational issues dominated early publications due to the challenges of rebuilding the Polish state, the field has evolved over time, with a growing share of articles addressing more diverse and complex themes. Using large language models for text classification, we demonstrate that 80-90% of the articles fit into the Ossowskis' five thematic categories, though a notable increase in unclassified articles in the 21st century suggests a broadening of SoS beyond its original conceptual framework.

Introduction

The Science of Science (SoS) as an academic discipline has a long and rich history, although for many researchers it remains an invisible part of science. Its origins date back more than 100 years, with the field primarily developing in Eastern Europe. The golden era of SoS occurred in the 1960s and 1970s, both in the East and the West. Contrary to the claims of Wang and Barabási in their book *The Science of Science* (Wang & Barabási, 2021), SoS is not an "emerging interdisciplinary field" driven by big data. Rather, it is part of a long-standing endeavor to study science through the tools of various disciplines, with philosophy, history, and sociology playing key roles. This is evident both in the early Western contributions to SoS, often associated with scholars like Jesmond D. Bernal, and in contemporary approaches to "research on research" or simply "metascience." (Krauss, 2024). The SoS programmatic foundations were rooted not in quantitative studies of science but in cultural, philosophical, and sociological understandings of science and its outcomes.

The aim of this paper is to demonstrate the disciplinary identity of a newly emerging academic discipline in Poland from 1918 to 2020. The year 1918 marks Poland's regaining of independence after 123 years and the founding of the world's first strictly science-of-science journal, *Nauka Polska. Jej Potrzeby, Organizacja i Rozwój* (in short: *Nauka Polska*, English: *Science and Letters in Poland: Their Needs, Organization, and Progress*). This journal continues to be published today, despite a 40-year interruption caused by the Sovietization of Poland's science and

higher education system. During this period, several Polish science-of-science journals were established, with two key ones being *Nauka (Polska)* [(Polish) Science] and *Zagadnienia Naukoznawstwa* [Problems of the Science of Science], both published by the Polish Academy of Sciences. These three journals served as platforms for discussion and publication by key SoS representatives from around the world, including Derek de Solla Price, Jesmond D. Bernal, Vasily Nalimov, and Gennady Dobrov.

To analyze the SoS disciplinary foundations, we use the classification of five areas of SoS presented by Maria and Stanisław Ossowski in the first programmatic article on the science of science. Utilizing large language models (LLMs), we will examine articles from the three aforementioned science-of-science journals from the years 1918 to 2020 to determine whether the classification proposed by the Ossowskis, based on experiences from the first two decades of the discipline's existence (up to 1935), remains useful for categorizing the SoS. For this purpose, we will use 9,272 full-text articles from three SoS journals from the *Corpus of Polish Science of Science Journals* (CPSSJ), which contains over 50,000 articles from 12 Polish science-of-science journals published between 1918 and 2020 (Kulczycki et al., 2023).

Polish origins of science of science

Why did the science of science emerge specifically in Poland in 1910-20s (Cain & Kleeberg, 2024; Kokowski, 2015)? The shortest possible answer is this: a group of Polish scholars, due to objective circumstances, was primarily educated outside the borders of what was then a non-existent Poland on the map of Europe. They participated in international research and discussions on the status and role of science. In terms of understanding what science is, how it should be practiced, and its role, there was nothing particularly unique to explain the emergence and development of the science of science in Poland. What was unique, however, at the turn of the 19th and 20th centuries, was the end of Poland's partition into three parts in 1918 after 123 years of non-independence.

In 1918, the Polish state was being built practically from scratch. There were no structures at the national level. The situation was also similar in science. Universities had already functioned in Poland for many centuries, but they were part of the science systems of the three states: Austro-Hungary, Germany and Russia. As part of the construction of the new structures of the state, work began on the consolidation of science. This challenge of unifying the three partitions and understanding the potential role of science in this task created the historical conditions for proposing the science of science in Poland. Scientists in no other country, who also discussed science and its role, faced the same political, cultural, and societal task as Polish scientists in similar historical circumstances.

In the Polish historiography of the science of science, it is accepted that three science of science programs emerged in Poland: Florian Znaniecki's in 1925, the Ossowskis' in 1935, and Kotarbiński's in 1965. However, it should be clarified that, although Znaniecki is considered a precursor of science of science programs, his foundational text, "Przedmiot i zadania nauki o wiedzy" [The Subject Matter and Tasks of the

Science of Knowledge] from (1925), is more of an encouragement to create a mature program (which the Ossowskis accomplished) than a mature program itself. In this work, Znaniecki first proposed the term "naukoznawstwo" (literally meaning *science connoisseurship* or *science studies*), whose equivalent in Polish is also the term "nauka o nauce" (science of science), proposed a decade later by the Ossowskis in 1935. In contrast, Kotarbiński's approach is best described as an analysis of the philosophical conditions for practicing the science of science.

In 20th-century Polish social sciences, the Ossowskis' name was one of the most prominent and influential. After World War II, both served as professors at the University of Warsaw, fulfilling key academic roles and playing significant social roles in resisting the Stalinization and Sovietization of Polish social and scientific life. A few years before the war, they co-authored a key text crucial to the development of the science of science. Thirty years after the publication of the Ossowskis' work, Bernal referred to their proposal during a congress in Warsaw in 1965, stating that the first use of the term *science of science* in its current sense should be attributed to the Ossowskis (Bernal & Mackay, 1966, p. 9).

Their 1935 article "Science of Science" is regarded as the first comprehensive and most important programmatic work in the field of science of science and we believe that it remains relevant and offers more than just historical value. It is worth noting that the text was available in English in *Organon* a year after its publication, though in limited circulation, later reprinted a couple of times, among others, in English in the 1960s in *Minerva* (Ossowska & Ossowski, 1964), the 1980s in a volume dedicated to the Polish contribution to science of science (Walentynowicz, 1982), and again in 2024 in a collective work on science of science in interwar Poland (Cain & Kleeberg, 2024). Despite these publications, the article is not widely known, even among contemporary scholars of science of science.

Five Areas of Science of Science

Since the publication of the Ossowskis' work, much has been written about science studies and the science of science itself. When re-reading their "Science of Science," it is crucial to remember that their approach—treating science as a social and cultural phenomenon—was far from obvious at the time. In fact, it was quite revolutionary during the 1920s and 1930s. The Ossowskis propose identifying five overlapping areas (as they note) that the science of science should study. They write that three are fundamental groups of science of science problems that would form the backbone of a new branch of science and add two areas of practical issues. According to the Ossowskis, the science of science consists of three fundamental groups of problems concerning *episteme*, the people of science, and the entire sector of science and higher education, along with their institutions. These three areas are:

1. *Philosophy of Science*, which considers, among other things, the concept of science (what it is and what it is not). This represents the epistemological perspective of the science of science.
2. *Psychology of Science*, which studies the mental development of the scientific worker.

3. *Sociology of Science*, which examines science in the context of social life and the entire cultural life. Within the sociology of science, the dependence of scientific development on economic conditions, the structure of a given society, and the organization of education are studied.

Additionally, the science of science encompasses two areas of practical problems:

4. *Practical-Organizational Issues*. The Ossowskis emphasize that research and reflection on these issues have thus far been primarily conducted by institutions dedicated to promoting science, which have applied theoretical results from the previously defined three areas to practical purposes. This area also includes science policy ("social and state policy towards science"). The Ossowskis note that this area deserves to be distinguished due to its practical nature.
5. *Historical Issues*. The study of the history of individual disciplines, the history of the researcher's concepts, and so forth, also has a practical dimension, as earlier mentioned areas or groups can utilize these studies in their work.

Materials and Methods

We analyzed 9,232 texts from three journals published between 1918 and 2020 (the last year included in the CPSSJ). Table 1 presents the quantitative characteristics of the journals' contents. We analyzed only articles published in Polish (the document count also includes editorial pages, tables of contents, announcements, and a few articles published in languages other than Polish).

Table 1. Characteristics of three analyzed journals.

Journal	Years	Documents	Articles
Nauka Polska. Jej Potrzeby Organizacja i Rozwój	1918–1920, 1923, 1925, 1927–1939, 1947, 1992-2020	1,516	1,095
Nauka (Polska)	1954–2020	7,844	6,024
Zagadnienia Naukoznawstwa	1965–2019	2,484	2,113
Total number of documents / articles		11,844	9,232

Each article from the three journals was stored as a text file. The mean text length was 32,785 characters, the median was 25,523 characters, and the 75th percentile was 43,288 characters. The longest article contained more than one million characters (it was a monographic issue on the history of an institution). To limit costs, the length of articles was capped at 80,000 characters. This truncation affected 52 articles, i.e., 5.56% of the texts.

Using the OpenAI platform (https://platform.openai.com), we prompted a Large Language Model, the GPT-4o, for each file using code written in Python. The GPT-4o model is a multilingual generative transformer developed by OpenAI and was released in May 2024. In total, GPT-4o was queried 9,232 times. The prompts for each article were independent of the others, so GPT-4o performed a full-text analysis

each time to assign the best category. The prompt was in Polish, and we asked GPT-4o to classify the article into one of five areas of the science of science indicated by the Ossowskis (as presented in the previous section) or to assign a 'non-classified category', if none of the five categories were appropriate. GPT-4o returned an answer for each article, including the assigned category and a justification. The category was extracted from the GPT-4o response using regular expressions. Both author of the study crosschecked the GPT-4o responses and agreed on the quality of the provided classification.

Results

We have analyzed the complete set of articles from the three journals included in our dataset, as well as the results for each journal separately. Figures 1 and 2 show the classification of articles across years and thematic areas. The highest number of articles was published during the so-called golden age of the Science of Science (SoS), in the 1970s. This peak can be attributed to both global and local factors. Globally, researchers across the world increasingly engaged with SoS themes, driven by Cold War-era research competition. Locally, Poland experienced a period of relative prosperity in the 1970s, which translated into greater availability of paper and the capacity to publish more extensive journal issues.

Figure 2 demonstrates that in the early years (1918–1939), practical-organizational issues dominated SoS publications. This focus was understandable, given the need to rebuild the Polish state and its science and higher education system after regaining independence. Over time, the prominence of this category declined, but it remained a dominant theme throughout the years. The analysis shows that 80-90% of the articles were successfully classified into one of the five categories proposed Ossowskis. However, starting from the 2000s, the percentage of unclassified articles approaches 20%, which may suggest that the conceptual scope of SoS has expanded beyond the original five areas. Confirming this hypothesis will require further, planned analyses.

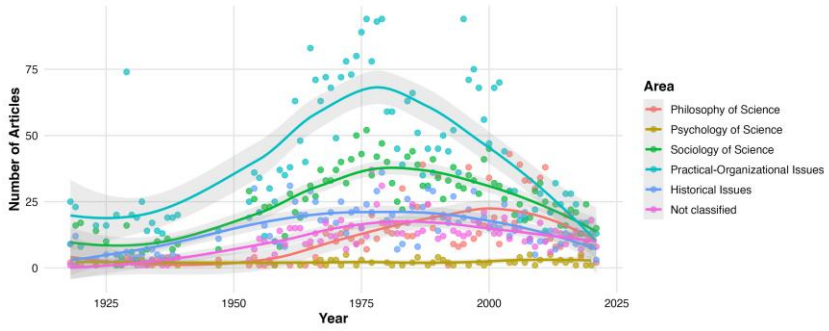


Figure 1. The number of articles per year across areas.

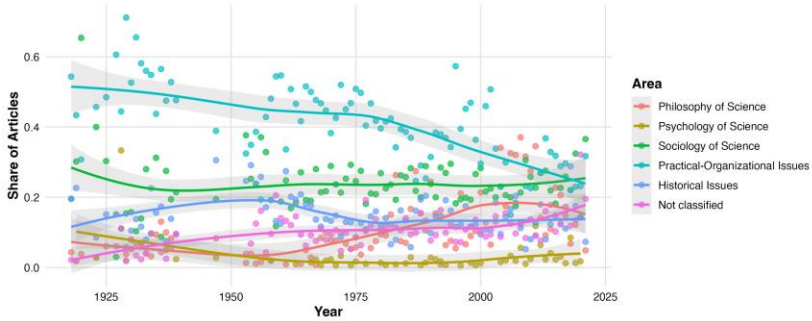


Figure 2. The share of articles across areas and years.

Figures 3 and 4 present the results broken down by individual journals. It is clear that the majority of articles were published in the main Polish Academy of ciences journal, Nauka (Polska), and primarily dealt with practical-organizational matters.

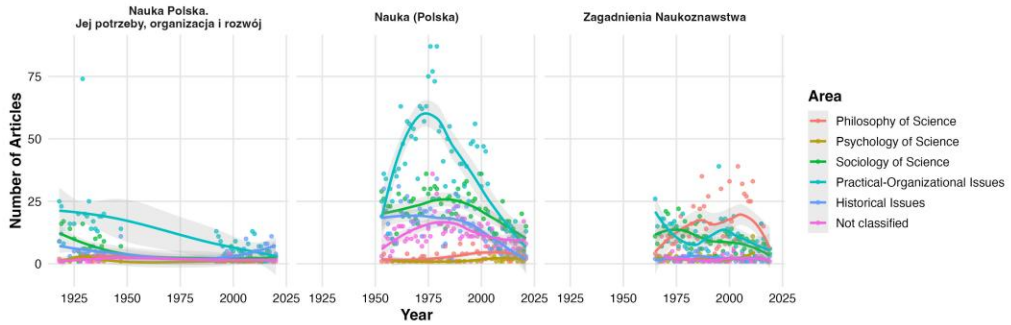


Figure 3. The number of articles per year across areas for each journal separately.

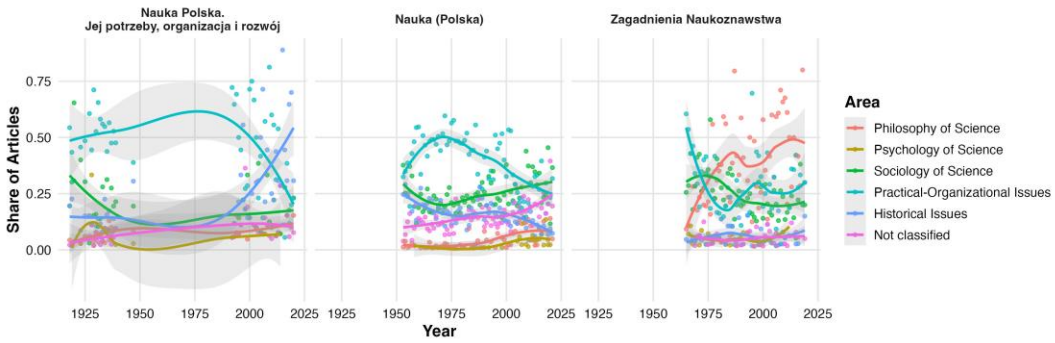


Figure 4. Share of articles across areas and years for each journal separately.

Interestingly, the profile of Zagadnienia Naukoznawstwa appears to be more philosophical-theoretical, aligning not only with the Polish approach to SoS but also with the emerging Soviet SoS, which emphasized a philosophical foundation rather than the sociological perspective more prevalent in the West during the same period (Aronova, 2011).

Conclusion and plans

This study highlights the distinct disciplinary identity and foundations of the SoS in Poland. Unlike the contemporary portrayal of SoS as a data-driven, emerging field, its roots in Poland reveal a well-established tradition that spans over a century. The key themes addressed in Polish SoS reflect both global intellectual trends and local historical circumstances, particularly the need to reconstruct the state and its scientific institutions following Poland's regaining of independence in 1918.

The results of our classification demonstrate that the thematic areas proposed by the Ossowskis remain relevant for understanding the historical trajectory of the science of science (SoS) in Poland. Some topics, particularly those related to the psychology of science, have been fading, even though they were crucial not only for SoS in the early 20th century but also for scientometrics (Godin, 2007). Moreover, the increasing share of unclassified articles in recent years indicates a diversification of approaches within the field. This evolution suggests that contemporary SoS is moving beyond the traditional framework, incorporating new methodologies and perspectives. Further research is needed to explore these developments and assess their implications for the field's future.

Our findings show the importance of recognizing the historical and cultural context in shaping the evolution of academic disciplines. The Polish case offers valuable insights into how SoS has been conceptualized and practiced in different geopolitical settings. Understanding these variations is essential for a more nuanced appreciation of the global history of science studies. The next phase of our research will involve extracting references from footnotes to analyze cited works. This will allow us to assess the extent to which the scientific discourse in Polish SoS journals has been localized, focusing predominantly on Polish authors and issues, versus its engagement with global scholarship. It will also enable an exploration of how this balance has shifted over the past century.

Acknowledgments

The work of PK was co-financed by the state budget under the program of the Ministry of Education and Science called 'Science for Society II' project no. NdS-II/SP/0460/2023/01 amount of co-financing 1 735 800 PLN total value of the project 1 735 800 PLN.

References

Aronova, E. (2011). The politics and contexts of Soviet science studies (Naukovedenie): Soviet philosophy of science at the crossroads. *Studies in East European Thought, 63*(3), 175–202. https://doi.org/10.1007/s11212-011-9146-y

Bernal, J. D., & Mackay, A. L. (1966). Towards a Science of Science. *Organon, 3*, 9–17.

Cain, F., & Kleeberg, B. (Eds.). (2024). *A New Organon: Science Studies in Interwar Poland* (1. Auflage, p. 550). Mohr Siebeck.

Godin, B. (2007). From Eugenics to Scientometrics: Galton, Cattell, and Men of Science. *Social Studies of Science, 37*(5), 691–728. https://doi.org/10.1177/0306312706075338

Kokowski, M. (2015). The Science of Science (Naukoznawstwo) in Poland: The Changing Theoretical Perspectives and Political Contexts – A Historical Sketch from the 1910s to 1993. *Organon, 47*, 147–237.

Krauss, A. (2024). *Science of Science: Understanding the Foundations and Limits of Science from an Interdisciplinary Perspective* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/9780198937401.001.0001

Kulczycki, E., Zambrano Mena, Y. A., & Krawczyk, F. (2023). Budowa i charakterystyka Korpusu Polskich Czasopism Naukoznawczych. *Zagadnienia Informacji Naukowej, 61*, 9–31.

Ossowska, M., & Ossowski, S. (1964). The science of science. *Minerva, 3*(1), 72–82.

Walentynowicz, B. (Ed.). (1982). *Polish contributions to the science of science*. PWN Polish Scientific Publishers.

Wang, D., & Barabási, A.-L. (2021). *The Science of Science* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108610834

Znaniecki, F. (1925). Przedmiot i zadania nauki o wiedzy. *Nauka Polska, 5*, 1–78.

https://doi.org/10.51408/issi2025_109

Distinguishing Types of Scientific Innovation Capacity: Exploring the Patterns and Dynamics of Knowledge Combinations and Impacts on Innovation in Biomedical Literature

Jinyu Gao[1], Yi Bu[2], Sarah Bratt[3]

[1]*jinyugao@arizona.edu*
College of Information Science, University of Arizona, 1103 E. 2nd St, Tucson, Arizona, 85721 (United States)

[2]*buyi@pku.edu.cn*
Department of Information Management, Peking University, 5 Yiheyuan Road, Haidian District, Beijing 100871 (China)

[3]*sebratt@arizona.edu*
College of Information Science, University of Arizona, 1103 E. 2nd St, Tucson, Arizona, 85721 (United States)

Abstract

Never-before-seen, groundbreaking ideas advance science, but so do combinations of ideas and prior knowledge. This paper identifies three types of scientific innovation capacities – digging, bridging, and jumping– based on three kinds of knowledge combinations: repeated, predicted, and unexpected combinations. The capacities and combinations are assessed by using concepts associated with papers in the biomedical literature (1950-2023) and link prediction methods. We analyzed concepts from the Semantic MEDLINE Database (SemMedDB) to understand how the combination of knowledge within national research systems reflects distinct innovation capabilities and, in turn, impacts national research performance. This paper has implications for scientific innovation policy and the quantitative study of networked concepts in biomedicine.

Introduction

Scientific innovation is often driven by the recombination of existing knowledge (Uzzi, Mukherjee, Stringer, & Jones, 2013). While previous studies have explored predictable and unpredictable combinations, these studies have largely overlooked repeated combinations, that is, combinations that reuse established links between concepts. This paper introduces a unified framework that classifies biomedical knowledge combinations into three types: repeated, predicted, and unexpected, corresponding to three forms of innovation capacity: digging, bridging, and jumping. Despite growing interest in how knowledge structures influence innovation, the relationship between different types of knowledge recombination and their specific roles in scientific advancement remains underexplored. In particular, few studies have considered all three combination types together, or examined how these patterns reflect and shape innovation capacity across both individual research outputs and national research systems. Using the large-scale semantic network SemMedDB and a link prediction method based on common neighbors, this study analyzed patterns of biomedical knowledge combinations. By examining how repeated,

predicted, and unexpected knowledge links are formed, the research aims to identify the role these combinations play in driving scientific innovation. The study also explores how these patterns vary across countries, providing insights into how different approaches to knowledge recombination reflect national differences in innovation capacity. This analysis will contribute to understanding how the structure of knowledge influences scientific progress and innovation outcomes on a global scale and has implications for scientific innovation policy and the quantitative study of networked concepts in biomedicine.

Related Studies

Combinatorial innovation

Understanding innovation has always been a key issue in the science of science, particularly in how to measure innovation and identify the factors that influence the innovation process. In early studies of innovation, Schumpeter (2003) argued that innovation is essentially a recombination of factors of production. Later studies came to show that recombination can, indeed, stimulate innovation. The way in which different types of knowledge are combined reflects distinct innovation patterns. For example, Uzzi, Mukherjee, Stringer, and Jones (2013) analyzed the combinations of references in scientific papers from the perspectives of atypicality and conventionality. They suggested that a low probability of two journals being cited together indicates novelty, while a high probability reflects conventionality. They found that the high impact papers stand on the shoulders of conventional and novel knowledge brought together. Veugelers and Wang (2019) further showed that scientific papers making rare journal combinations are more likely to be cited by patents. This suggests a direct technological impact. Such papers are also more likely to be cited by other papers with high technological impact. Another perspective on combinatorial innovation lies in disruptiveness and consolidation. Scientific reward is also coupled with risk. As such, scientists must manage the trade-off between consolidation and disruptiveness in scientific innovation. Studies have also used a later-published papers' citation behavior to a focal paper and its references as a strategy of evaluating the disruptiveness of a paper. For a focal paper and its references, there has three different citation strategies for a future paper: 1) cited the reference(s) of the focal paper but not the focal paper, 2) cited the focal paper and its reference(s) together, 3) cited the focal paper only without any of its references, and the innovation extent of the focal paper increase from the consolidation of tradition to disruptive (Funk & Owen-Smith, 2017; Wu, Wang, & Evans, 2019). Ample studies have analyzed innovation and novelty from the perspective of recombination based on network structure. Foster, Rzhetsky, and Evans (2015) analyzed how chemical knowledge is combined in scientific research. They identified five research strategies: new consolidation, new bridge, repeat consolidation, repeat bridge, and jump. These strategies are based on whether scientists connect two chemical entities within the same research area (clustering) and whether the study involves new chemicals. Their results showed that risky innovations – those focused on new knowledge or novel relationships – can lead to greater impact than stable innovations

built on established knowledge and relationships. Hofstra et al. (2020) introduced two types of novelty: conceptual novelty, which measures the number of knowledge concept pairs linked for the first time in a thesis abstract, and impactful novelty, which refers to how often these novel combinations are used in future theses. They found that gender and racial minorities tend to produce more innovative and semantically distant combinations. However, these novel contributions receive less adoption. The study revealed that it is more difficult for underrepresented groups to maintain their academic positions.

Predicting research trends

The rapid surge in the volume of scientific literature presents a significant challenge for researchers. As a result, many studies have started exploring methods for predicting research trends. For example, Shi, Foster, and Evans (2015) constructed hypergraphs to connect authors, chemicals, diseases, and methods within each paper. The chemicals, diseases, and methods were extracted from MeSH (Medical Subject Headings). The results revealed that the network distance in the biomedical hypergraphs was relatively small, with most new links forming between nodes that were already neighbors or only two steps apart. Krenn and Zeilinger (2020) built a co-occurrence network from quantum physics papers and used neural networks for link prediction to predict research trends. Their findings revealed that emerging concepts and new connections can be related to key discoveries and advancements in quantum science. Shi and Evans (2023) found that unexpectedly novel combinations of article keywords (MeSH terms, PACS codes, USPC codes) and cited journals tend to be associated with high-impact papers, ranking in the top 10% by citation count.

Unequal scientific development among countries/geographic regions

In recent years, some studies have begun to analyze the national innovation capacity of countries. Studies demonstrate marked inequalities in national scientific development. For example, Miao et al. (2022) used revealed comparative advantage (RCA) to analyze national scientific development, treating disciplines as “products” of nations. They identified three discipline clusters linked to economic advantages, showing that while nations diversify research, global science is increasingly specialized. The study highlighted inequalities, especially in low-income countries, and called for policies to bridge disparities and build scientific capacity. Gomez, Herman, and Parigi (2022) proposed "the citation well" to assess citation distortion by comparing international citation flow and publication similarity. They used QAP network regression to show how core countries are over-cited while peripheral ones are under-cited, revealing how unequal knowledge recognition hinders national scientific development.

Data and Methods

Datasets

SemMedDB: The Semantic MEDLINE Database (Kilicoglu, Shin, Fiszman, Rosemblat, & Rindflesch, 2012) is a repository of semantic triples (subject CUIs –

predicate – object CUIs) extracted from PubMed, where CUIs refers to Concept Unique Identifiers in the Metathesaurus which is belong to Unified Medical Language System (UMLS) (Bodenreider, 2004).

PubMed Knowledge Graph (PKG) 2.0: PKG 2.0 is a comprehensive knowledge graph dataset integrating over 36 million papers, 1.3 million patents, and 0.48 million clinical trials in biomedicine (Xu et al., 2024). The country information for each paper is determined based on the first affiliation of the first author.

Link prediction

The SemMedDB dyads (subject CUIs – object CUIs) are used to build the undirected and unweighted network $G(V, E)$, where V is the set of nodes and E is the set of links. Prediction network is denoted as $G_y \epsilon [t-w, t)$, while the focal network is denoted as $G_y = t$, where t refers to focal year and w represents the time window. The time window used in this paper is 5 years. Edges that will be linked together in the future are predicted based on the concept of common neighbors. A common neighbor is a node that connects to both of two other nodes, and having more of these shared connections means those two nodes are more likely to be linked in the future. (Lü & Zhou, 2011). The common neighbor edges satisfy the following conditions:

1) Not present in the prediction network $G_y \epsilon [t-w, t)$: The edge (u, v) or (v, u) does not exist in prediction edges, ensuring that the selected edges are potential new edges.

2) Nodes share common neighbors: There is at least one common neighbors between nodes u and v.

Preliminary Results

This section presents the main findings, illustrated through four figures. Each figure highlights a different aspect of the analysis, covering the distribution of edge types, combinations of innovation capacities, and their effects on citation and influence. The following results provide a detailed look at these patterns.

Figure 1 (top) illustrates the growth in the number of papers recorded in SemMedDB, showing a clear upward trend from 1950 to 2023. The number of CUIs studied each year has also increased, although at a slower pace compared to the number of papers. The gray bars indicate the number of new biomedical concepts introduced each year, relative to all previous data. There was a sharp increase in new CUIs in 1975, followed by another significant surge in 2020.

Figure 1 (bottom) shows the overlap of CUIs between papers published each year and those published in the previous year, the past three years, and the past five years. In both 1975 and 2020, the overlap decreased due to the influx of new CUIs. Nevertheless, the overlap with CUIs from the past five years remained high, consistently ranging from 80% to 90%.

Figure 2 (top) illustrates how edges in the focal year's network are classified using a link prediction method. Potential links identified by the common neighbor method are termed “common neighbor edges.” If such edges are realized in the focal network, they are classified as “predicted edges.” Edges overlapping with those from previous networks are labeled as “repeated edges.” The focal network may also contain

"unexpected edges," including those between existing nodes with no common neighbors, between new and existing nodes, or between two new nodes. Figure 2 (bottom) presents the proportions of the three edge types from 2000 to 2023. Repeated edges dominate, accounting for about 70%, followed by predicted edges (20–30%), while unexpected edges remain below 10%.

This paper focuses on the years 2000, 2010, and 2020 to examine the distribution of three edge types and their corresponding innovation capacities. Repeated edges represent *digging* innovation capacity, predicted edges indicate *bridging* capacity, and unexpected edges reflect *jumping* capacity. These capacities represent specific types of innovation capacity. Figure 3 presents a ternary plot (top) and a stacked bar chart (bottom), illustrating the portfolios of the three edge types across countries. The analysis shows that most countries rely heavily on repeated combinations, supplemented by predicted ones, while unexpected combinations are relatively rare. Figure 4 (top) illustrates that each paper can contain multiple edges, and the combinations of these edges form different edge combination types. These combinations include the mix of predicted and repeated edges (type id = 1), papers with only repeated edges (type id = 6), and those with only predicted edges (type id = 7). Additionally, papers can contain all three types of edges (type id = 2), combinations of unexpected and repeated edges (type id = 4), papers with only unexpected edges (type id = 5), and combinations of unexpected and predicted edges (type id = 3). Figure 4 (bottom) suggests that the combination of different types of innovation capacities leads to varying impacts on a paper's citation and influence.

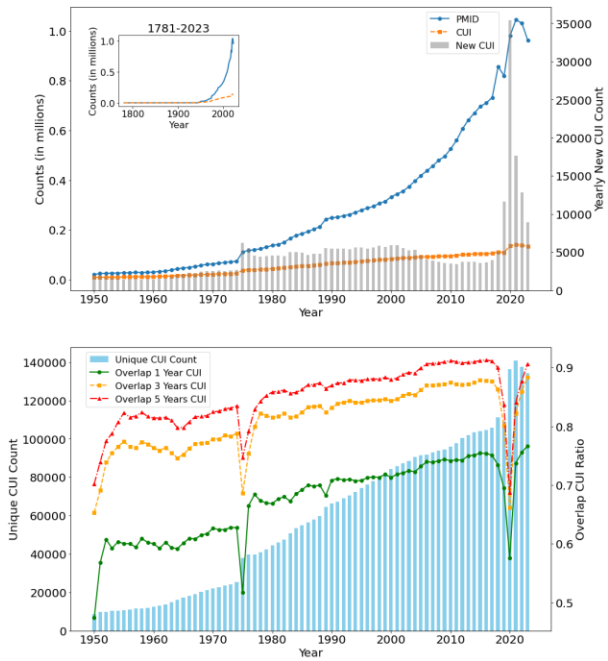


Figure 1. Summary of PMID, CUIs, and Yearly New CUIs in SemMedDB (1950-2023).

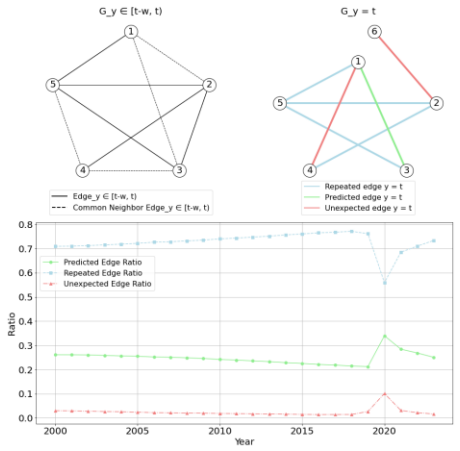


Figure 2. Edge types in concept graph: repeated, predicted, and unexpected.

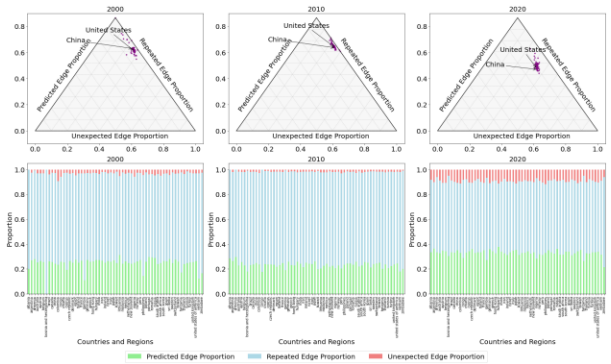


Figure 3. Distribution of Edge Types among Countries.

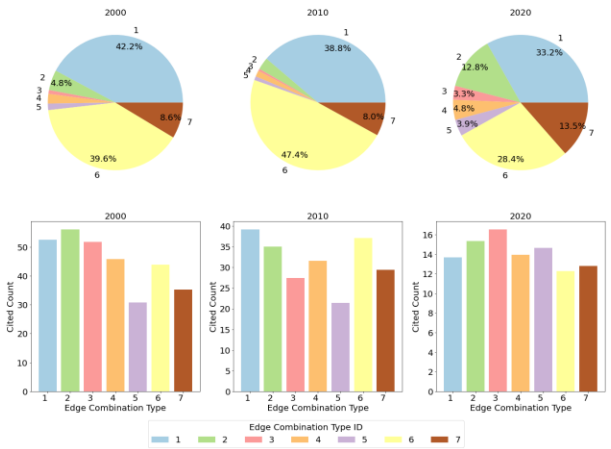


Figure 4. Edge Combination Types.

Conclusions

This paper explores different types of research innovation capacities by analyzing knowledge combinations based on biomedical entities utilizing a link prediction method through common neighbors. The knowledge combinations are divided into three types: repeated edges, predicted edges, and unexpected edges, corresponding to digging, bridging, and jumping innovation capacities, respectively. The advantage of identifying innovation capacity portfolios at both the national and paper levels is it reveals that scientific research relies heavily on repeated edges and predictable links. These predictable edges have at least one common neighbor, and their proximity within the network is crucial for advancing scientific development. Several areas remain open for improvement. First, this study focuses on dyads extracted from triples, overlooking the relational context between nodes. Future research could use triples to construct knowledge graphs and better leverage their richer semantic information. Second, the classification of edges could be further refined, for example, the unexpected edges might be distinguished based on whether they involve newly introduced biomedical entities. Additionally, both predicted and unexpected edges represent new connections. Analyzing their likelihood of being adopted in future research would provide valuable insights into the dynamics of scientific innovation and the diffusion of novel ideas. Finally, the innovation capacity portfolio and its correlation with scientific recognition could be more accurately analyzed through regression or even causal inference techniques in the future.

References

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(90001), 267D – 270.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. American Sociological Review, 80(5), 875–908. SAGE Publications Inc.

Funk, R. J., & Owen-Smith, J. (2017). A Dynamic Network Measure of Technological Change. Management Science, 63(3), 791–817.

Gomez, C. J., Herman, A. C., & Parigi, P. (2022). Leading countries in global science increasingly receive more citations than other countries doing similar research. Nature Human Behaviour, 6(7), 919–929. Nature Publishing Group.

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The Diversity–Innovation Paradox in Science. Proceedings of the National Academy of Sciences, 117(17), 9284–9291. Proceedings of the National Academy of Sciences.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: A PubMed-scale repository of biomedical semantic predications. Bioinformatics, 28(23), 3158–3160.

Krenn, M., & Zeilinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. Proceedings of the National Academy of Sciences, 117(4), 1910–1916. Proceedings of the National Academy of Sciences.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 390(6), 1150–1170.
Miao, L., Murray, D., Jung, W.-S., Larivière, V., Sugimoto, C. R., & Ahn, Y.-Y. (2022). The latent structure of global scientific development. Nature Human Behaviour, 6(9), 1206–1217. Nature Publishing Group.
Schumpeter, J., & Backhaus, U. (2003). The Theory of Economic Development. In J. Backhaus (Ed.), Joseph Alois Schumpeter: Entrepreneurship, Style and Vision (pp. 61–116). Boston, MA: Springer US. Retrieved April 24, 2025, from https://doi.org/10.1007/0-306-48082-4_3
Shi, F., & Evans, J. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Nature Communications, 14(1), 1641. Nature Publishing Group.
Shi, F., Foster, J. G., & Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. Social Networks, 43, 73–85.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. Science, 342(6157), 468–472. American Association for the Advancement of Science.
Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. Research Policy, 48(6), 1362–1372.
Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. Nature, 566(7744), 378–382. Nature Publishing Group.
Xu, J., Yu, C., Xu, J., Ding, Y., Torvik, V. I., Kang, J., Sung, M., et al. (2024, October 10). PubMed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. arXiv. Retrieved January 2, 2025, from http://arxiv.org/abs/2410.07969

https://doi.org/10.51408/issi2025_122

Distribution Differences of Knowledge Diversity among Authors in Different Contributor Roles—Evidence from 101014 PLOS ONE Articles

Jingyuan Li [1], Yi Zhao [2], Jiaqi Zeng [3], Chengzhi Zhang [4]

[1] *lijingyuan2002@njust.edu.cn*, [2] *yizhao93@njust.edu.cn*, [3] *zengjq@njust.edu.cn*, [4] *zhangcz@njust.edu.cn*

Department of Information Management, Nanjing University of Science and Technology, 210094 Nanjing (China)

Abstract

In the fast-developing academic environment, the composition and structure of research teams are becoming more diversified and complex. Authors with different roles in the team also show obvious differences in knowledge diversity. Understanding of these differences not only helps to dissect the laws of academic development, but also effectively promotes individual career development and teamwork. Therefore, based on 101,014 papers published in PLOS ONE (2017 – 2023), author knowledge diversity is calculated using pre-publication academic outputs from the OpenAlex dataset. Additionally, we explore the distribution patterns of knowledge diversity among authors in different research roles. The results of the study show that organizational roles such as Funding Acquisition are more likely to be undertaken by academics with a high degree of knowledge diversity. Technical roles such as Data Curation and Investigation can be finished by authors with relatively lower knowledge diversity. In addition, the study reveals gender differences in knowledge diversity and role taking. Male authors focus on overall design role and female authors are more involved in experiment. This study not only provides a strong empirical basis for the promotion of interdisciplinary collaboration and the development of innovation ability, but also provides a new theoretical perspective for a deeper understanding of the career development of researchers.

Introduction

With the rapid development of current scientific research, single-discipline knowledge become inadequate to solve the complex scientific challenges (Guimerà et al., 2005). Multidisciplinary knowledge reserve has become essential, offering foundational bases and novel perspectives for scientific research. Authors' knowledge diversity, or their interdisciplinary knowledge reserves, significantly impacts their ability to deal with complex problems. Consequently, knowledge diversity has emerged as a key metric for evaluating authors' learning and innovation capabilities.

Knowledge diversity measures the engagement breadth of authors across different disciplines (Chang, 2012). However, current research primarily focuses on team-level knowledge diversity, which fails to accurately capture individual diversity. Existing metrics, such as Rao-Stirling index (Stirling, 2007), explore the link between team interdisciplinarity and research impact. Yet, these indicators emphasize differences among team members rather than individual knowledge diversity across disciplines. Moreover, they rely on post-publication data analysis, resulting in a lag in information acquisition (Zheng et al., 2022).

The roles authors assume in research teams reflect their specific contributions to a project. With the standardization of author contribution statements, the investment of authors in knowledge, skills and labour can be more precisely quantified, providing new ways of thinking about analyzing their actual role contributions (Clement, 2014). Authors' knowledge diversity is closely tied to their roles in research programs, as individuals with varying levels of diversity tend to take on different roles and make distinct contributions (Yang et al., 2022). Consequently, knowledge diversity across roles may exhibit significant differences.

In summary, standardized author contribution statements enable the study of roles and labor division within research teams. While progress has been made in analyzing team-level knowledge diversity, research on individual author knowledge diversity and its distribution across roles remains limited. This gap hinders a deeper understanding of team knowledge structures and the enhancement of research efficiency and innovation. To address this, our study calculates author knowledge diversity using data from PLOS ONE journals (2017−2023) and the OpenAlex platform. Besides, we explore how knowledge diversity is distributed among authors in different roles within research teams.

Related Work

Knowledge Diversity in Research Teams

Research teams are usually organised in terms of outputs, and all authors of a paper are considered as a whole (Zhang & Guo, 2019). Team knowledge diversity can be divided into team shared knowledge diversity and individual author knowledge diversity in the team. The former focuses on the overall knowledge composition of the team, while the latter focuses on the degree of cross-disciplinary of individual members (Chang, 2012). The current research mainly focuses on the knowledge diversity at the team level, while less attention is paid to the knowledge diversity of individual authors in the team. For example, Chowdhary et al. (2024) found that knowledge diversity in enduring collaborative teams has a positive influence on productivity but a negative influence on its impact. Zheng et al. (2022) showed that teams with high expertise diversity do not have a significant effect on their impact in the short term but attract more interdisciplinary citations in the long term. Zhang and Guo (2019) argued that knowledge diversity has a double-edged effect in cross-functional teams. Knowledge leaders can modulate its impact on team performance through the interactive memory system.

Role and Contribution of authors

Scientific collaborations increasingly favour multi-authorship, with a declining proportion of sole-authored papers (Wuchty et al.,2007). Contributions usually refer to the division of labour among co-authors (Rahman et al., 2020), while roles reflect the specific contributions of authors in scientific research. Therefore, clarifying roles and contributions is crucial for improving research efficiency and quality (Yang et al., 2022). Earlier studies measured contribution based on the order of attribution,

with the value of contribution decreasing with the order of attribution (Das & Das, 2020). However, studies have found that in some fields, the first and last authors contribute more and the middle authors contribute less (Sundling, 2023). In order to clarify the contribution of authors, many journals use classification systems (Larivière et al., 2020). Among them, some of the journals under the PLOS initially used a five-role taxonomy before fully introducing the Contributor Role Taxonomy (CRediT) in 2016(McNutt et al., 2018). Based on this, Li et al. (2023) developed mapping schemes to analyze the differences in the distribution of author contributions under different systems. Macaluso et al. (2016) found that females were more inclined to experimental work, while males were more likely to take on other roles. These studies highlight the complexity of role division in research teams and offer new insights into understanding author contributions.

Data and Methodology

Data collection and pre-processing

PLOS ONE [1] is an international, peer-reviewed journal that publishes multidisciplinary and interdisciplinary research. Since 2016, it has adopted the CRediT system, a 14-category role framework. This study uses OpenAlex[2] dataset, a global knowledge graph, which provides real-time, multi-dimensional academic data through algorithms and data mining. Additionally, Genderize.io, a widely recognized gender identification tool based on author names, is used to determine gender accurately. This study selects data from PLOS ONE papers from 2017 to 2023, which is conducted with OpenAlex dataset and the Genderize.io[3] tool.

Data collection and processing involve four steps. Firstly, we collected and preprocess metadata and contribution statements from PLOS ONE. Secondly, we extracted author contributions using two formats: line-break-separated text (requiring rule-based abbreviation matching) and JSON text (directly parsed). Both methods ensure accurate mapping of authors to their contributions. Thirdly, author publication counts and concept scores were retrieved to calculate knowledge diversity by utilizing DOIs to connect to OpenAlex. Finally, Genderize.io is utilized to determine author gender, while unidentifiable data is excluded. The final dataset comprises 101,014 articles and 405,766 authors from PLOS ONE journals.

After pre-processing the data, knowledge diversity trends are analyzed, and gender differences are compared. Additionally, role participation rates and gender disparities are examined using contribution statements. Finally, the percentage of authors in each role type within specific diversity intervals is analysed.

Measurement of the authors' knowledge diversity

Knowledge diversity measures the interdisciplinary scope of authors. A lower value indicates a more focused field, while a higher value reflects broader disciplinary involvement and balanced expertise.

[1] https://journals.plos.org/plosone/
[2] https://openalex.org/
[3] https://genderize.io/

OpenAlex defines 19 core top-disciplines. It predicts the topics to which papers belonged from information such as their titles and abstracts, assigning concept score (0-1) of the 19 disciplines for each paper (Priem et al., 2022). For this study, author annual knowledge diversity is calaulated by a 19-dimensional vector. And each dimension reflects the average concept scores of their pre-year papers in each discipline. After normalizing, knowledge diversity is quantified using Equation 1:

$$KD_j = -\sum_{i=1}^{19} p(x_{ij}) \log_2 p(x_{ij}) \quad (1)$$

Where KD_j denotes the author knowledge diversity in year j; *and* $p(x_{ij})$ denotes the normalised value of the concept score in subject i in year j. To ensure comparable results, the final normalisation was done again using $log_2(19)$. A value of 0 indicates single-topic focus, while 1 represents a balanced knowledge structure across all disciplines.

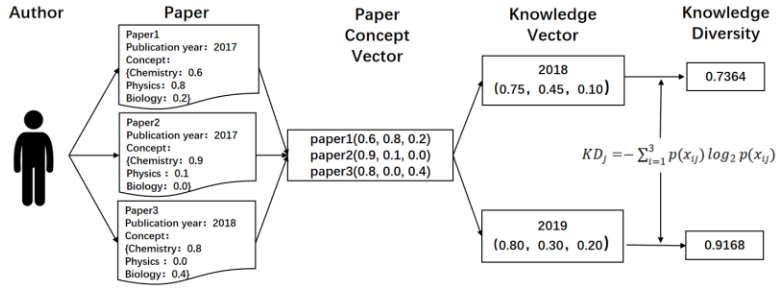


Figure 1. Example of knowledge diversity calculation.

For example, an author published three papers. The publication year and concept scores provided by OpenAlex for Chemistry, Physics and Biology are presented in Fig.1. Firstly, we construct paper concept vectors for paper1, paper2 and paper3.The vector values represent the concept scores of the three disciplines given by OpenAlex. Secondly, we calculate the annual knowledge vectors of the author. The values of each dimension of the vector represent the average concept scores of all papers published by the author before that year in each discipline. For example, the score in Chemistry in 2018 is the average of the concept scores in Chemistry of paper1 and paper2, i.e., (0.6+0.9)/2=0.75, and the same for other disciplines, which ultimately leads to the knowledge vector of the author in 2018 as (0.75, 0.45, 0.10). After normalisation, we calculated its knowledge diversity in 2018 as 0.7364 using Equation 1. Similarly, the knowledge vector and knowledge diversity in 2019 can be calculated using the concept scores of paper1, paper2 and paper3(Fig.1).

Result

Trends in knowledge diversity and gender differences

As shown in Figure 2, the average annual knowledge diversity of authors remains stable, ranging between 0.215 and 0.260. From 2017 to 2022, knowledge diversity shows minimal fluctuation but rises significantly from 2022 to 2023, peaking in 2023. This increase may be driven by the growing use of tools like large models,

which have broadened research horizons and enhanced interdisciplinary knowledge integration. For example, large models in medicine have boosted transfer learning, interdisciplinary collaboration, and educational training. It allows authors to integrate multi-disciplinary expertise (Karabacak & Margetis, 2023). From a gender perspective, men's knowledge diversity is significantly higher than women's in every year, although the gap narrows between 2021 and 2023. This may be influenced by the fact that female academics are, on average, younger than their male counterparts (McChesney & Bichsel, 2020).

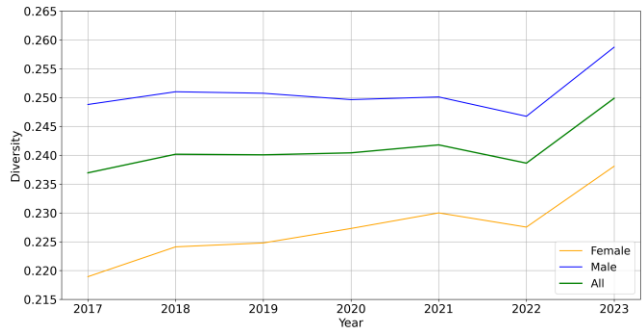


Figure 2. Average annual distribution of knowledge diversity.

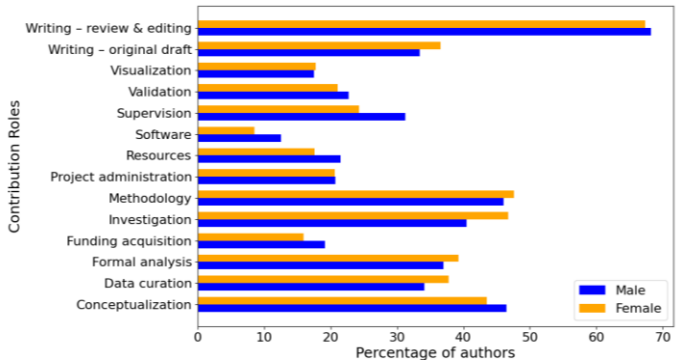


Figure 3. Gender share of roles by contribution type.

Frequency of authors' participation in roles and gender differences

We examine gender differences in research roles by calculating the participation rates of male and female authors in each role (Fig. 3). The results reveal significant variations in role participation frequencies. Writing - review & editing is the most common role, with over 60% participation for both genders, while Software had the lowest, at less than 15%. In terms of gender differences, male are more often involved in conceptual tasks such as Funding acquisition and Supervision. In contrast, female are more often involved in experiment roles such as Investigation, Data curation (Larivière et al., 2020).

Differences in the distribution of knowledge diversity among authors in different roles

The overall distribution of knowledge diversity ranges from 0.0 to 0.7 and its main part is concentrated in the interval of 0.15 to 0.4. Figure 4 depicts the distribution of various roles on the knowledge diversity dimension. It can be observed that most of the roles exhibit a high frequency distribution in the interval of medium knowledge diversity (0.2-0.4), while the frequency in the interval of high knowledge diversity (0.6-0.7) is extremely low, with a frequency close to 0. Particularly noteworthy are the frequency peaks in the intersection of certain roles with knowledge diversity zones, which are significantly higher than in other zones. Funding Acquisition, for example, has a higher distribution of knowledge diversity in the medium-high range than the other roles. It suggests that this role is more likely to be taken on by members with a broader knowledge background. And it is generally performed by leaders with deeper and broader knowledge (Chinchilla-Rodríguez et al., 2019). In other practice-specific roles like Data Curation and Investigation, authors with relatively low knowledge diversity can still perform the work. This suggests these tasks rely less on broad knowledge and more on deep expertise in a specialized area.

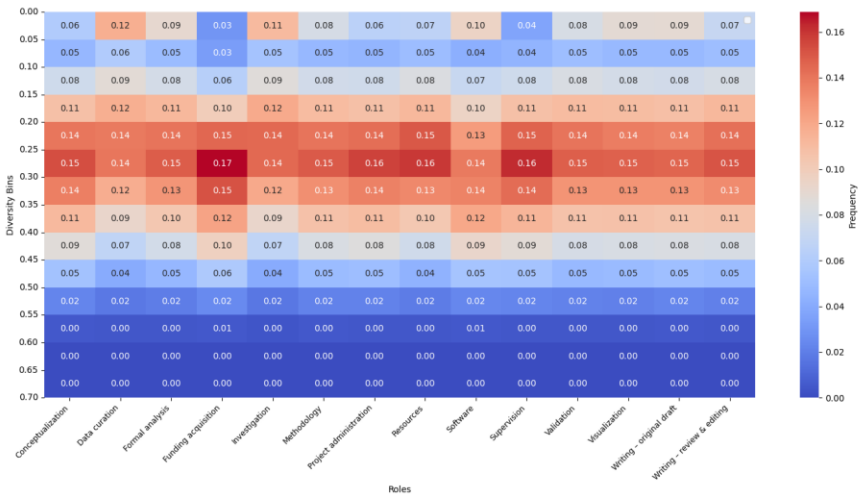


Figure 4. The frequency of distribution of knowledge diversity across contributing roles.

Conclusion

This study analyzes data from PLOS ONE journals (2017–2023), revealing gender differences in knowledge diversity and role distribution. Male authors tend to engage more in conceptual roles, while female authors are more involved in experiment-related roles. Over time, the gap in knowledge diversity between genders narrowed. Additionally, roles like Funding Acquisition require higher knowledge diversity, whereas technical roles (e.g., Data Curation, Investigation) need lower requirements. Despite these insights, the study has limitations. The data, limited to PLOS ONE

journals (2017–2023), may lack generalizability despite the journal's interdisciplinary scope. Future research could expand to other journals and extend the time frame to validate findings. Advanced statistical methods, like causal and correlation analysis, can better examine the link between knowledge diversity and role contributions. This provides refined insights for optimizing research team structures and improving scientific efficiency.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Grant No.72074113).

References

Chang, W. J. (2012). Differential effects of knowledge diversity on team innovation: An agent-based modeling. 2012 International Conference on Innovation Management and Technology Research. https://doi.org/10.1109/ICIMTR.2012.6236384

Chinchilla-Rodríguez, Z., Sugimoto, C. R., & Larivière, V. (2019). Follow the leader: On the relationship between leadership and scholarly impact in international collaborations. PLOS ONE, 14 (6), e0218309. https://doi.org/10.1371/journal.pone.0218309

Chowdhary, S., Gallo, L., & Musciotto, F. (2024). Team careers in science: Formation, composition and success of persistent collaborations. arXiv preprint. https://doi.org/10.48550/arXiv.2407.09326

Clement, T. P. (2014). Authorship matrix: A rational approach to quantify individual contributions and responsibilities in multi-author scientific articles. Science and Engineering Ethics, 20 (2), 345-361. https://doi.org/10.1007/s11948-013-9454-3

Das, N., & Das, S. (2020). 'Author contribution details' and not 'authorship sequence' as a merit to determine credit: A need to relook at the current Indian practice. The National Medical Journal of India, 33 (1), 24-27. https://doi.org/10.4103/0970-258X.308238

Guimerà, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. Science, 308 (5722), 697-702. https://doi.org/10.1126/science.1106340

Karabacak, M., & Margetis, K. (2023). Embracing large language models for medical applications: Opportunities and challenges. Cureus, 15 (3), e39305. https://doi.org/10.7759/cureus.39305

Larivière,V., Pontille, D. , & Sugimoto, C. R. . (2020). Investigating the division of scientific labor using the contributor roles taxonomy (credit). Quantitative Science Studies, 2(9070), 1-24. https://doi.org/10.1162/qss_a_00097

Li, K., Zhang, C., & Larivière, V. (2023). Are research contributions assigned differently under the two contributorship classification systems in PLOS ONE? arXiv preprint. https://doi.org/10.48550/arXiv.2310.11687

Macaluso, B., Larivière, V., & Sugimoto, T. (2016). Is science built on the shoulders of women? A study of gender differences in contributorship. Academic Medicine, 91 (8), 1136-1142. https://doi.org/10.1097/ACM.0000000000001261

McChesney, J., & Bichsel, J. (2020). The aging of tenure-track faculty in higher education: Implications for succession and diversity. Knoxville, TN: College and University Professional Association for Human Resources. https://doi.org/10.13140/RG.2.2.18555.95521

McNutt, M., Bradford, M., & Drazen, J. (2018). Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. OSF Preprints. https://doi.org/10.31219/osf.io/asywp

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint. https://doi.org/10.48550/arXiv.2205.01833

Rahman, M. T., Regenstein, J. M., & Kassim, N. L. A. (2020). Contribution based author categorization to calculate author performance index. Accountability in Research, 27 (1), 1-20. https://doi.org/10.1080/08989621.2020.1860764

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. Journal of The Royal Society Interface, 4 (15), 707-719. https://doi.org/10.1098/rsif.2007.0213

Sundling, P. (2023). Author contributions and allocation of authorship credit: Testing the validity of different counting methods in the field of chemical biology. Scientometrics, 128 (3), 2737-2762. https://doi.org/10.1007/s11192-023-04680-y

Wuchty, Stefan, Jones, Benjamin, F., & Uzzi, et al. (2007). The increasing dominance of teams in production of knowledge. Science. https://doi.org/10.1126/science.1136099

Yang, S., Xiao, A., & Nie, Y. (2022). Measuring coauthors' credit in medicine field - Based on author contribution statement and citation context analysis. Information Processing & Management, 59 (3), 102924. https://doi.org/10.1016/j.ipm.2022.102924

Zhang, L., & Guo, H. (2019). Enabling knowledge diversity to benefit cross-functional project teams: Joint roles of knowledge leadership and transactive memory system. Information and Management, 56 (8), 103156. https://doi.org/10.1016/j.im.2019.03.001

Zheng, H., Li, W., & Wang, D. (2022). Expertise diversity of teams predicts originality and long-term impact in science and technology. arXiv preprint. https://doi.org/10.48550/arXiv.2210.04422

https://doi.org/10.51408/issi2025_093

Does Distance Still Matter? The Impact of Geographic Patterns in Scientific Knowledge Sourcing on Invention Value

Guiyan Ou[1], Chaocheng He[2], Jiang Wu[3]

[1]*Ouguiyan@whu.edu.cn*
Wuhan University, School of Information Management, Wuhan (China)

[2]*he_chaocheng@whu.edu.cn*
Wuhan University, School of Information Management, Wuhan (China)
Wuhan Unversity Shenzhen Research Institute, Shenzhen, Guangdong (China)

[3]*jiangw@whu.edu.cn*
Wuhan University, School of Information Management, Wuhan (China)

Abstract

While the "death of distance" hypothesis suggests that advanced information technologies have diminished the role of geography in knowledge flows, the question "Does distance still matter?" remains debated in scientific knowledge utilization. This study examines the influence of geographic factors on invention value by analyzing the spatial patterns of scientific knowledge sources cited in patents. Using a sample of 463,393 science-based patents granted by the USPTO, we investigate three geographic dimensions: geographical distance, external knowledge proportion, and geographical diversity. Our analysis demonstrates that the geographical distance of scientific knowledge sources negatively impacts both the technical and economic value of inventions. Furthermore, we find curvilinear (inverted U-shaped) relationships between invention technical value and both the proportion of external scientific knowledge and geographical diversity of knowledge sources. Notably, higher proportions of external scientific knowledge demonstrate a significant negative association with the economic value of patented inventions. These results challenge the "death of distance" hypothesis and demonstrate that geographical patterns remain crucial in scientific knowledge utilization. Our findings advance the understanding of geography's role in science-based innovation and offer important implications for strategic knowledge sourcing in the digital age.

Introduction

While technological advancement has become increasingly dependent on scientific discovery, the translation of scientific innovation into technological innovation remains a complex and uncertain endeavor. Early empirical evidence demonstrates that the diffusion of ideas, knowledge, and innovations is fundamentally constrained by geographical factors (von Graevenitz et al., 2022). And the geographical factors can largely shape the landscape of scientific as well as technological innovation. However, extant literature has paid limited attention to examining how the geographic patterns of scientific knowledge sources may influence the efficacy of transforming scientific discoveries into technological innovations.

The "death of distance" hypothesis, emerging from the rapid advancement of information and communication technologies, suggests that geographical constraints on knowledge flows have diminished in the digital age. This hypothesis posits that modern technological infrastructure has created a "placeless" paradigm of knowledge dissemination, theoretically enabling efficient access to globally distributed scientific knowledge (Abramo et al., 2020). However, recent empirical evidence challenges this notion, indicating that while digital technologies have reduced some geographic barriers, the influence of spatial distance on knowledge flows remains substantial (von Graevenitz et al., 2022). This persistent role of geography in knowledge transfer raises a fundamental question that motivates our research: Does geographic distance still matter in the transformation of scientific knowledge into valuable technological innovations?

Scientific knowledge, characterized by its codified nature and open standards, readily transcends geographic boundaries (Wang, 2024). This knowledge can be classified as local or external based on its spatial acquisition patterns across various geographical scales, from municipal to national boundaries. Overreliance on local knowledge sources, while convenient, risks cognitive lock-in and redundancy, potentially hindering innovation in increasingly complex systems (Hohberger and Wilden, 2022). Thus, understanding how varying intensities of external scientific knowledge input shape technological innovation outcomes becomes crucial.

Moreover, extant research demonstrates that geographical diversity in knowledge acquisition significantly influences innovation performance (Subramaniam and Yound, 2005). For instance, Belderbos et al. (2018) suggested that geographical diversification in Corporate Venture Capital (CVC) portfolios directly enhances firm-level innovation performance. Yet, the complexity of processing knowledge from diverse contexts imposes substantial coordination and integration costs (Lahiri, 2010; Singh, 2008). Consequently, the impact of geographical diversity in scientific knowledge sources on technological innovation outcomes merits rigorous investigation.

Technological inventions, representing the initial realization of novel concepts and principles, form the foundational basis of technical innovation. The value of these inventions serves as the critical metric for innovation success, fundamentally reflecting innovation effectiveness. This study examines science-based innovation by focusing on the geographical patterns of scientific knowledge sources. Specifically, we address three critical questions:

（1）How does the geographical distance of scientific knowledge sourcing affect invention value?

（2）What is the impact of external scientific knowledge proportion on invention value?

（3）How does the geographical diversity of scientific knowledge sources influence invention value?

Data and Variables

Data Sources

This study examines the science-technology linkages by analyzing patent citations to scientific literature. The empirical analysis draws on two primary data sources. First, we utilize patents granted by the United States Patent and Trademark Office (USPTO) from 2001 to 2010. Second, we employ the "Reliance on Science" (RoS) dataset (Marx and Fuegi, 2020), which contains 42,822,458 citation links between patents (both US and non-US) and scientific publications indexed in Microsoft Academic Graph (MAG). Our final sample consists of 463,393 USPTO granted patents that cite at least one MAG-indexed publication during the 2001-2010 period. These patents from 6,472,102 unique patent-to-publication citation pairs, linking to 1,407,439 distinct scientific publications in the MAG database.

Variables

We assessed invention value through two distinct dimensions: technological impact and economic value. We operationalized technological impact through the count of forward citations received within five years of patent grant (*Citation_5*) (Hsu et al., 2021). For economic value, we employed a binary measure of patent assignment (*Is_Assignment*) (Kwon,2020), coded as 1 if the patent was transferred and 0 otherwise.

We investigated three independent variables: geographical distance (*Geo_Dist*), the proportion of external scientific knowledge (*Exter_Know_Ratio*), and geographical diversity (*Geo_Dive*). The first independent variable—*Geo_Dist*—captures the spatial separation between scientific paper authors and patent inventors. Following Gao and Rai's (2023) approach, we employed the Haversine formula for *Geo_Dist* calculation, which is specified as:

$$Distance_{a,b} = 2r \cdot \arcsin\left(\sqrt{sin^2\left(\frac{\varphi_a - \varphi_b}{2}\right) + \cos(\varphi_a)\cos(\varphi_b)\, sin^2\left(\frac{\lambda_{a-} - \lambda_b}{2}\right)}\right)$$

where φ_a and φ_b denote the latitudes of a and b, respectively, while λ_a and λ_b represent their corresponding longitudes. The constant r represents Earth's radius, which is approximated as 6,371 kilometers.

Exter_Know_Ratio is measured as the ratio of cited scientific papers originating from countries different from the inventor's home country. This metric is calculated as follows:

$$Exter_Know_Ratio = \frac{Num_Paper_{nonlocal}}{Num_Paper}$$

Where Num_Paper represents the total number of scientific papers cited by the patent, and $Num_Paper_{nonlocal}$ refers to the count of cited papers from countries other than the inventor's home country. *Geo_Dive* captures the spatial distribution of cited scientific papers in patents, which we quantify using the Blau index:

$$Geo_diversity = 1 - \sum_{i=1}^{C} p_i^2$$

where C denotes the total number of countries from which cited papers originate, and p represents the proportion of papers from each country. The resulting *Geo_Dive* measure ranges from 0 to 1, with higher values indicating greater geographical dispersion of scientific knowledge sources cited in the patent.

Our analysis accounts for multiple factors that potentially influence invention value (Zhu et al., 2022; Büttner et al., 2022; Poege et al., 2019), including PCT patent, patent scope, technological maturity, technological diversity, technological originality, number of inventors, number of applicants, backward patent citations, scientific publication count, scientific maturity, and impact of scientific publications. Additionally, we include dummy variables controlling for applicant organization type, application year, and technological field.

Given the count nature of Citations_5 and the binary structure of Is_Assignment, we employ negative binomial regression and logistic regression models for estimation, respectively.

Results

Table 1 summarizes our regression analyses examining the impact of three key variables (*Geo_Dist, Exter_Know_Ratio and Geo_Dive*) on invention value. The empirical results from Models 1 and 4 demonstrate that *Geo_Dist* has significant negative effects on both *Citations_5* (β=-0.039, p<0.001) and *Is_Assignment* (β=-0.041, p<0.001). These findings suggest that greater geographical distances in scientofic knowledge absorption are associated with reduced forward citations and lower probability of commercial transactions for patents.

Models 2 and 5 demonstrate significant negative coefficients for both linear and quadratic terms of *Exter_Know_Ratio* in predicting *Citations_5* and *Is_Assignment*. While this pattern initially suggests potential inverted U-shaped relationships, subsequent graphical analyses yield divergent findings. Figure 1(a) confirms a robust inverted U-shaped relationship between *Exter_Know_Ratio* and *Citations_5*, indicating an optimal level of external knowledge integration for maximizing citation impact. However, Figure 1(b) reveals that the estimated inflection point (-1.636) lies outside the observed data range, invalidating the hypothesized curvilinear relationship for *Is_Assignment*. Instead, *Exter_Know_Ratio* exhibits a consistently negative linear relationship with assignment probability, suggesting that increased

external knowledge integration systematically reduces commercial transfer likelihood.

Model 3 demonstrates significant negative coefficients for both the linear and quadratic terms in the relationship between Geo_Dive and Citations_5. The empirical visualization in Figure 1(c) corroborates this inverted U-shaped relationship, suggesting that while moderate levels of geographical diversity enhance patent citation performance, excessive spatial distribution introduces coordination challenges that ultimately impede innovation. Contrary to expectations, we found no significant relationship between the Exter_Know_Ratio and Is_Assignment.

Table 1. The Impact of Knowledge Sourcing Geographical Patterns on Invention Value.

Variables	Citations_5			Is_Assignment		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Geo_Dist	-0.039***			-0.041***		
Exter_Know_Ratio		-0.096***			-0.052***	
Exter_Know_Ratio 2		-0.044***			-0.016***	
Geo_Dive			-0.020***			0.009
Geo_Dive2			-0.039***			-0.003
PCT	-0.299***	-0.259***	-0.297***	-0.122***	-0.108***	-0.132***
Ln (Patent_Scope)	0.248***	0.244***	0.248***	0.086***	0.087***	0.089***
Ln (Tech_Maturity)	-0.179***	-0.179***	-0.181***	-0.155***	-0.154***	-0.156***
Ln (Tech_Diversity)	-0.130***	-0.124***	-0.127***	-0.084***	-0.084***	-0.086***
Tech_Originality	-0.142***	-0.139***	-0.145***	0.074***	0.069***	0.069***

Ln (Num_Inventors)	0.080***	0.080***	0.081***	0.057***	0.054***	0.054***
Ln (Num_Applicants)	-0.103***	-0.089***	-0.099***	1.121***	1.119***	1.115***
Ln (Num_PR)	0.396***	0.390***	0.396***	0.150***	0.148***	0.151***
Ln (Num_NPR)	-0.019***	-0.032***	-0.008***	0.051***	0.043***	0.044***
Ln (Sci_Maturity)	-0.134***	-0.129***	-0.130***	-0.083***	-0.081***	-0.082***
Ln (Sci_Citation)	0.040***	0.036***	0.040***	0.001	0.002	0.004*
OrgTypes	Yes	Yes	Yes	Yes	Yes	Yes
Time	Yes	Yes	Yes	Yes	Yes	Yes
IPC	Yes	Yes	Yes	Yes	Yes	Yes
_cons	2.758***	2.828***	2.778***	0.182***	0.202***	0.187***
N	353205	369799	369799	353205	369799	369799

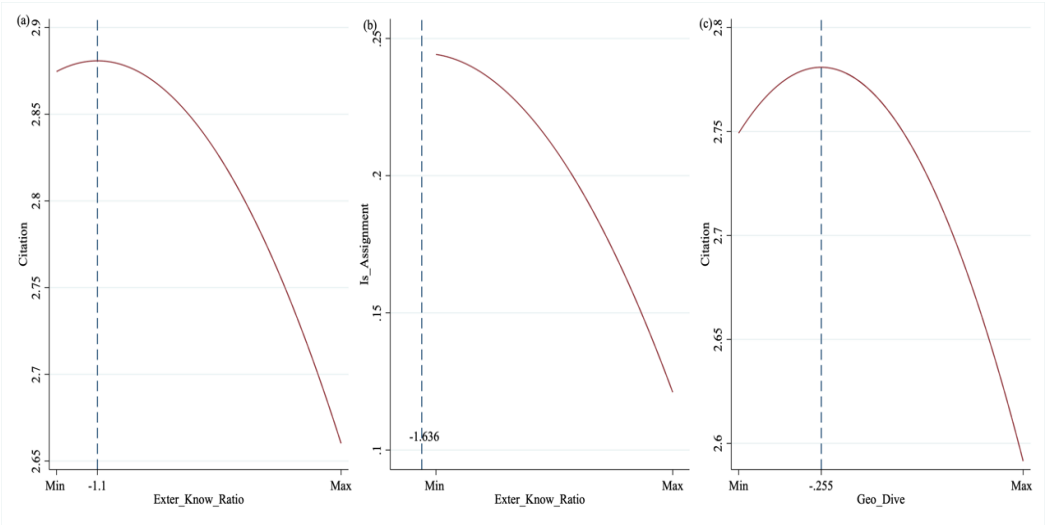


Figure 1. Nonlinear Effects of *Exter_Know_Ratio* and *Geo_Dive* on Invention Value with Identified Threshold Points (dashed lines).

Conclusion

This study advances our understanding of how geographical patterns of scientific knowledge sources influence invention value by examining three critical dimensions: geographical distance, external knowledge proportion, and geographical diversity. Through a systematic analysis of USPTO patent records integrated with RoS patent-science citation data, our empirical investigation yields several significant insights:

First, the geographical distance of scientific knowledge sources exhibits negative effects on both technological and economic value of patent inventions. While the advent of internet technologies has substantially reduced communication and interaction costs, potentially eliminating geographical barriers to knowledge dissemination in certain domains, our findings demonstrate that spatial distance remains a significant impediment in technological innovation. This persistent distance effect likely stems from the challenges in acquiring tacit knowledge necessary for effectively utilizing the codified knowledge embedded in scientific publications.

Second, we identify a nuanced relationship between external scientific knowledge proportion and invention value. Specifically, there exists an "optimal range" for external knowledge proportion in relation to technological value, while it demonstrates a consistently negative impact on economic value. Within this optimal range, increased input of external scientific knowledge correlates positively with patent citations. However, beyond a certain threshold, higher proportions of external scientific knowledge become detrimental to innovation quality. From a commercial perspective, greater reliance on external scientific knowledge in the technological development process corresponds to decreased likelihood of patent transfer.

Finally, our findings reveal an inverted U-shaped relationship between geographical diversity and technological value of patents, while showing no significant association with economic value. Patent forward citations increase with geographical diversity up to an optimal threshold, beyond which additional diversity becomes detrimental to technological impact.

This study has certain limitations. The temporal scope of our analysis (2001-2010) was deliberately chosen to accommodate patent transfer windows. Future research opportunities exist in extending this analysis to more recent periods by adopting economic value indicators that require shorter observation windows. Additionally, our analysis of patents across all technological domains does not account for varying degrees of science-dependency across different fields, potentially masking field-specific patterns.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (72204189), Guangdong Basic and Applied Basic Research Foundation (2022A1515110972) and Digital Intelligence Humanities Fundation of Wuhan University (2024SZWK023)

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2020). Does the geographic proximity effect on knowledge spillovers vary across research fields?. Scientometrics, 123, 1021-1036.

Belderbos, R., Jacob, J., & Lokshin, B. (2018). Corporate venture capital (CVC) investments and technological performance: Geographic diversity and the interplay with technology alliances. *Journal of Business Venturing*, *33*(1), 20-34.

Büttner, B., Firat, M., & Raiteri, E. (2022). Patents and knowledge diffusion: The impact of machine translation. *Research Policy*, *51*(10), 104584.

Gao, X., & Rai, V. (2023). Knowledge acquisition and innovation quality: The moderating role of geographical characteristics of technology. *Technovation*, *125*, 102766.

Hohberger, J., & Wilden, R. (2022). Geographic diversity of knowledge inputs: The importance of aligning locations of knowledge inputs and inventors. *Journal of Business Research*, *145*, 705-719.

Hsu, D. H., Hsu, P. H., Zhou, T., & Ziedonis, A. A. (2021). Benchmarking US university patent value and commercialization efforts: A new approach. *Research Policy*, *50*(1), 104076.

Kwon, S. (2020). How does patent transfer affect innovation of firms?. *Technological Forecasting and Social Change*, *154*, 119959.

Lahiri, N. (2010). Geographic distribution of R&D activity: how does it affect innovation quality?. *Academy of management journal*, *53*(5), 1194-1209.

Poege, F., Harhoff, D., Gaessler, F., & Baruffaldi, S. (2019). Science quality and the value of inventions. *Science advances*, *5*(12), eaay7323.

Singh, J. (2008). Distributed R&D, cross-regional knowledge integration and quality of innovative output. *Research Policy*, *37*(1), 77-96.

Subramaniam, M., & Youndt, M. A. (2005). The influence of intellectual capital on the types of innovative capabilities. *Academy of Management journal*, *48*(3), 450-463.

von Graevenitz, G., Graham, S. J., & Myers, A. F. (2022). Distance (still) hampers diffusion of innovations. *Regional Studies*, *56*(2), 227-241.

Wang, F. (2024). Does the recombination of distant scientific knowledge generate valuable inventions? An analysis of pharmaceutical patents. *Technovation, 130*, 102947.

Zhu, K., Malhotra, S., & Li, Y. (2022). Technological diversity of patent applications and decision pendency. *Research Policy*, *51*(1), 104364.

https://doi.org/10.51408/issi2025_104

Enhancing Scientometrics Prediction under Uncertainty: A DIKW-Based Framework and Methodological Synthesis

Shuya Chen[1], Guo Chen[2]

[1]996512152@qq.com,[2]delphi1987@qq.com

Nanjing University of Science and Technology, No. 200 Xiao Ling Wei, Nanjing, Jiangsu (China)

Abstract

This paper analyzes the uncertainties present in predictive-oriented scientometric research and, through a literature review, organizes and categorizes information analysis tasks related to prediction under uncertain conditions. Furthermore, to better adapt to these tasks, we approach the issue from the perspective of the DIKW model and summarize various methods for handling uncertainty. Finally, we propose a research framework for conducting predictive-oriented scientometric studies in uncertain environments, using scenario analysis and signal analysis to dealing with uncertainty.

Introduction

The advancement of information technology and the continuous progress of globalization have led to an exponential growth of open-source information. Its multi-source, complex, abundant, and uncertain nature has become the norm in modern information environment. Such an information environment has prompted profound changes in information analysis tasks, gradually shifting from targeted services with clear objectives to innovative and foresight-oriented information services in an environment filled with uncertainty (Zhao & Zeng, 2022). Alongside the increase in open-source information, uncertainties in scientometric research have also become more pronounced (Zhao, 2022). Although the widespread application of machine learning and artificial intelligence has driven the transformation of scientometrics toward a model-based analytical paradigm, such quantitative analysis methods ignore the uncertainties that are intrinsic to prediction problems. However, the incomplete and complex nature of information inevitably leads to uncertainties in predicting future trends. Existing scientometric analysis methods are unable to predict and evaluate the direction of future changes and their ramifications while ignoring the issue of uncertainty (Sun & Ke, 2007). Although the uncertainties inherent in prediction tasks, arising from various factors, cannot be entirely eliminated, scientific methods can be employed to significantly reduce uncertainties in scientometric analysis to the greatest extent possible (Wu et al., 2022).

Methods

This study employs literature review and content analysis methods, focusing on

informetric tasks related to prediction under uncertain environments and approaches for handling uncertainty. The scope of the review primarily centers on the field of library and information science, encompassing various literature resources such as academic journal articles, professional books and government reports. During the retrieval process, multiple authoritative databases were utilized. Chinese literature was mainly sourced from CNKI (China National Knowledge Infrastructure), while English literature was obtained from databases such as Web of Science, ScienceDirect, and SpringerLink. Initially, a simple search query was constructed using core concepts and related terms, such as ("uncertainty environment" OR "uncertainty" OR "uncertainty handling") AND "prediction". Following preliminary research, the search terms were expanded to include synonyms. For example, "uncertainty environment" was expanded to include "complex environment", and "prediction" was extended to "foresight" and "early warning". Additionally, sentence-level searches were conducted to ensure that relevant literature lacking specific keywords was not omitted. Table 1 lists the keywords and filters used in the two rounds of retrieval. We systematically reviewed the research content, methods, and conclusions of the selected literature, analyzing their contributions to handling uncertainty. The screened literature was then categorized and summarized to provide a structured overview of the issue.

Table 1. Search Items and Filters.

Search Items		Filters
Initial Search Items	**Expanded Search Items**	
Uncertainty	Fuzzy, Rough, Deep uncertainty	• Library and Information Science •Journal Articles OR book OR Report OR Conference Paper • Citations > 0
Uncertainty environment	Complex Environment	
Prediction	Early Warning, Forecast	
Uncertainty Handling	Uncertainty Representation, Uncertainty Measurement	

Results

Uncertainty in Predictive-Oriented Scientometrics

Uncertainty is an inherent challenge in scientometrics. Throughout the entire analytical process, from data input and analysis to the generation of results, various forms of uncertainty are always present. The goal of informetric research is to

leverage available information to reduce uncertainty in understanding and predicting phenomena, making prediction an essential information service. (Li & Sun, 2024). In predictive-oriented scientometric research, uncertainty manifests in three dimensions: information, process, and outcomes, as shown in Figure 1.

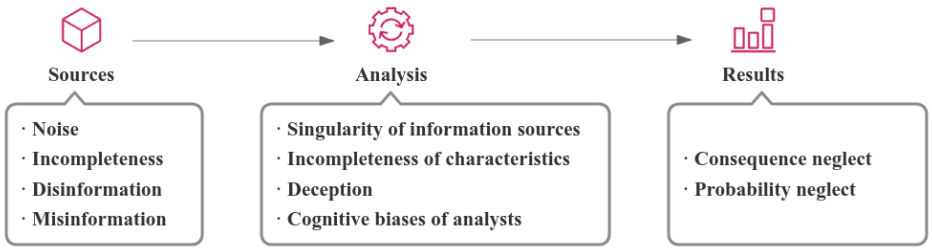


Figure 1. Uncertainty of Predictive-Oriented Scientometrics.

The Signal-Noise Theory suggests that an imbalanced signal-to-noise ratio is a significant source of uncertainty in intelligence analysis. Chinese scholar Wang Yanfei proposed the concept of "information fog", highlighting the falsehood and incompleteness of information. During the information analysis process, factors such as single-sided information sources, incomplete features, the deception, and cognitive biases of analysts can all contribute to uncertainty (Chen et al., 2022).The uncertainty issues in predictive intelligence outcomes include Consequence Neglect and Probability Neglect (Friedman & Zeckhauser, 2012). Consequence Neglect refers to overemphasizing the probabilities of various outcomes while neglecting their potential consequences. Probability Neglect refers to presenting possible outcomes without paying attention to the probabilities associated with each outcome.

Trend extrapolation is commonly used for prediction in informetric studies. To further illustrate why predictive-oriented scientometrics methods need to consider uncertainty, we use trend extrapolation as an example. Traditional data-driven scientometrics methods rely on quantitative analysis, aiming to identify the most likely patterns, thereby enhancing the certainty of a particular outcome and replacing all possibilities with the highest probability. For instance, as shown in Figure 2, curve l2 provides the best fit at time T1. However, at time T2, curve l3 becomes the best fit. During the transition from T1 to T2, changes in the phenomenon may be influenced by new factors or may no longer be affected by previously relevant factors.

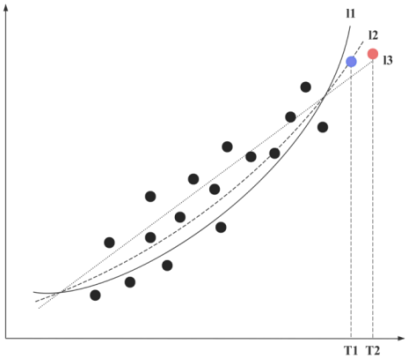


Figure 2. Future Trend Prediction Based on Historical Data.

Woodrow J. Kuhns (2003) argued that prediction methods should not simply produce results but also describe trends and the factors or variables influencing the development of a situation. Proposing future scenarios can lay a solid foundation for decision-making. Traditional trend extrapolation methods are limited by their overemphasis on identifying the best methods and outcomes. The issue lies in the fact that almost all scientometrics research is based on incomplete information, and the quantity and quality of information can both contribute to uncertainty in predictive-oriented scientometrics research (Mandel, 2020). Our goal is not to find a method that can completely eradicate uncertainty, but rather to develop one that can further minimize uncertainty in the analysis process and go beyond static and simplistic conclusions.

Predictive-Oriented Tasks under Uncertain Environments

In the context of big data, the explosive growth of open-source information has profoundly influenced scientometric research. While deterministic information has become increasingly accessible, this very accessibility underscores the importance of analyzing uncertain information. Concurrently, the intricate and layered nature of information uncertainty has not only amplified the demand for predictive capabilities but also imposed more rigorous standards on predictive tasks. Based on a comprehensive review and synthesis of the literature, we have delineated seven key predictive-oriented scientometric tasks in complex environments. These tasks are systematically classified based on their role in the information chain within the DIKW (Data-Information-Knowledge-Wisdom) model, the degree of uncertainty they involve, and their strategic significance, as illustrated in Figure 3.

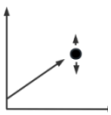
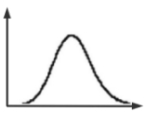
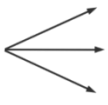
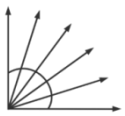
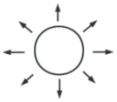
The DIKW model, rooted in Ackoff's classification of human cognitive content, distinguishes four levels: Data, Information, Knowledge, and Wisdom (Bosancic, 2016). Data consist of symbols representing the attributes of things or objects,

which are inherently devoid of meaning. To convert data into information, it is essential to collect, organize, and process data relevant to the target problem, extract meaningful components, and contextualize them. Transitioning from fragmented information to systematic and theoretical knowledge requires extensive induction, analysis, and synthesis. Knowledge encompasses theories and patterns derived by individuals, while wisdom involves applying knowledge to solve problems and make decisions. In uncertain environments, rather than striving for optimal decisions, the emphasis shifts to flexible and nuanced decision-making that can adapt to multiple future scenarios.

The second dimension is strategic significance. From the perspective of information analysis, "strategy" focuses on the "information activities of the subject". Without a subject, there is no drive or uniqueness in competition and confrontation, because information activities of the subject are closely connected to its economic, social, and cultural background (Yang, 2022). The closer the analysis is to the tactical level, the finer the granularity of the problems, such as resource replenishment and information fusion. Conversely, the higher the strategic significance, the more macroscopic the problems to be considered, requiring the mobilization and coordination of resources across various domains, as seen in tasks like early warning.

The third dimension is the degree of information uncertainty. Marchau (2019) categorized uncertainty into four levels based on the nature of the problems, as shown in Table 2. Across all levels of uncertainty, resource replenishment remains a necessary task. Information fusion, however, is relevant except in scenarios with fully determined objectives and information. When multiple foreseeable futures exist but their likelihoods are uncertain, tasks such as technology foresight and intelligence assessment become essential. Level 4 uncertainty, the deepest level, can be divided into two scenarios: one where the future is constrained by many plausible possibilities (4a) and another where we only know that we do not know (4b). Tasks such as clue discovery, situational awareness, counterintelligence, and early warning aim at addressing this profound level of uncertainty, requiring not only the exploration of possibilities based on existing knowledge but also forward-looking expertise and insights.

Table 2. Levels of Information Uncertainty.

Degree of Uncertainty	*Level 0*	*Level 1*	*Level 2*	*Level 3*	*Level 4*	
Objectives	Complete Determinism	A clear enough future	Alternate futures with probabilities	A few plausible futures	Many plausible futures	Unknown future
						

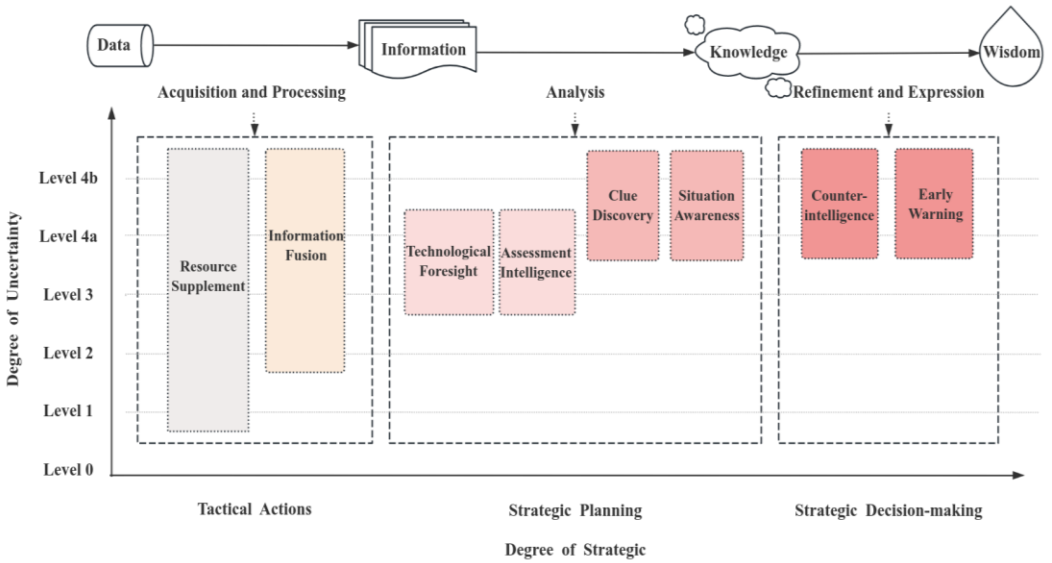


Figure 3. Predictive-Oriented Tasks and Their Classification in Uncertain Environments.

Scientometrics Methods for Addressing Uncertainty

Information uncertainty has become increasingly prominent in complex information environments, and the focus of information analysis has gradually shifted from descriptive intelligence to predictive, evaluative, and early-warning intelligence. These tasks in uncertain environments, as illustrated in Figure 3, often rely on traditional data-driven methods, which are ill-suited for tasks aimed at

reducing future uncertainty. Predictive work in uncertain environments emphasizes the extensive collection of information, the handling of knowledge uncertainty, and the acknowledgment of multiple future possibilities to support decision-making.

In past scientometric research, numerous methods have been developed to address various uncertainties, including uncertainties in information sources, uncertainty relationship mining, and uncertainty representation. Building on the DIKW model, we categorize the collected methods into three types based on the uncertainties present at each stage of cognitive content transformation. As shown in Figure 4, uncertainties such as information representation exist during the transition from Data to Information. To better quantify uncertainty and represent uncertain information, we need to use foundational uncertainty representation methods to store and express information and knowledge in a more scientific form, making them applicable to complex uncertain scenarios. During the transition from Information to Knowledge, the core interest is identifying which "signals" or knowledge can help us predict future events or capture potentially unnoticed information in uncertain environments. Ultimately, during the transition from Knowledge to Wisdom, we need to synthesize, analyze, and organize knowledge at a macro level to generate insightful content to support decision-making across diverse issues. This method collection, constructed based on the characteristics of each stage of the DIKW model, covers various uncertainty issues that may arise throughout the entire knowledge chain, from raw data collection to the generation of high-value intelligence.

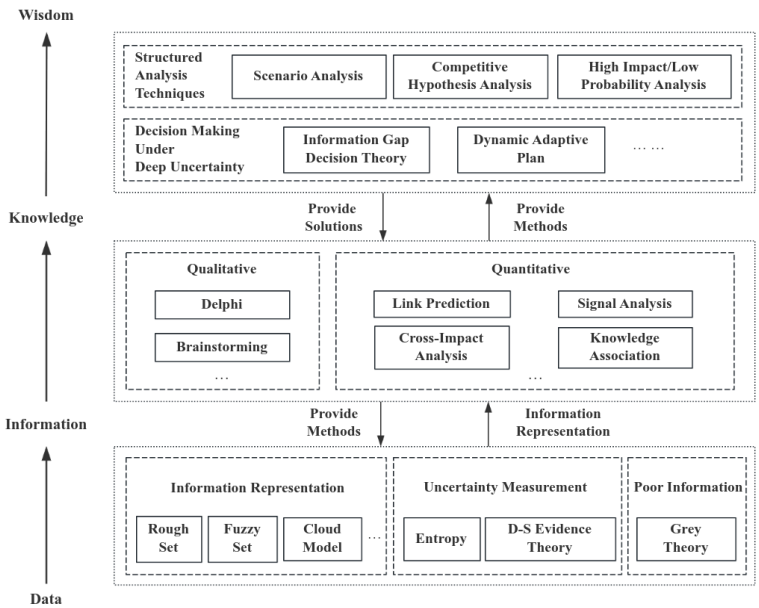


Figure 4. Method Framework for Scientometrics in Uncertain Scenarios.

Toward New Scientometric Approaches: A Case Study of Signal Analysis

In previous research, we have revealed the widespread presence of uncertainty in predictive tasks from a scientometric perspective and explored a collection of methods to address these uncertainties. Future research will further focus on optimizing the combination of methods at the practical level, proposing to integrate scenario analysis with weak signal analysis to construct a novel analytical framework for technology trend foresight, as shown in Figure 5.

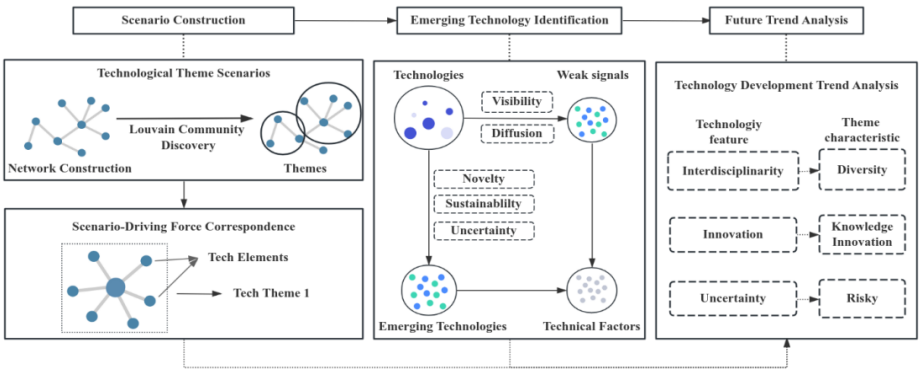


Figure 5. Practical Framework for Technology Foresight in Uncertain Environments.

Specifically, due to the inherently high uncertainty of technological evolution, especially in long-term trend foresight, where potential influencing factors are complex and intertwined, traditional single-metric methods struggle to comprehensively capture their dynamic characteristics. Thus, this study proposes to incorporate weak signal analysis by identifying and filtering technologies with potential influence as driving forces, combined with scientometric methods (such as network analysis and topic modeling) to quantitatively analyze the evolutionary paths of technological themes. To concretely reveal the trends of future scenarios, we set three core predictive objectives: diversity, innovation, and risk, each scenario characteristic, corresponding scientometric indicators will be established. Through driving force analysis and scientometrics analysis, scenario-based predictions will be made for trend foresight of each technological theme.

Acknowledgement

This work is supported by Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX24_0791).

References

Bosancic B. (2016). Information in the knowledge acquisition process. *Journal of Documentation,* 72(5), 930-60.

Chen Y., Ma X. J., et al. (2022). Studies on Analytical Method System of Warning Intelligence: From a Perspective of Probability Analysis. *Journal of Intelligence*, 41(02), 20-28+11.

Friedman J. A., (2012). Zeckhauser R. Assessing Uncertainty in Intelligence. *Intelligence and National Security*, 27(6), 824-847.

Kuhns W. J. (2003). *Paradoxes of Strategic Intelligence*. London: Routledge.

Li Y. & Sun J. J. (2024). Predictive Intelligence Research Based on Data Intelligence: A New Paradigm for Predictive Intelligence Research in a New Environment. *Journal of the China Society for Scientific and Technical Information*, 43(6), 633-643.

Mandel D.R. (2020). Assessment and Communication of Uncertainty in Intelligence to Support Decision Making: Final Report of Research Task Group SAS-114.

Marchau V. A. W. J., Walker W. E., et al. (Ed.). (2019). Decision Making under Deep Uncertainty: From Theory to Practice. Cham: Springer.

Sun J. J. & Ke Q. (2007). A Method for Information Analyses under Incomplete Information Environment: Scenarios Analysis and Its Application in Information Research. *Library and Information Service*, 51(02), 63-66+120.

Wu Y., Ma J., et al. (2022). Uncertainty in Intelligence Analysis: A Study from the Perspective of National Security Intelligence. *Information studies: Theory & Application*, 45(05), 66-74.

Yang G. L. (2022). Development of Intelligence Studies under the National Strategies. *Journal of the China Society for Scientific and Technical Information*, 41(07), 762-773.

Zhao Z. Y. & Zeng W. (2022). Constructing a Scientific and Technological Intelligence Theoretical System in a Complex Information Environment. *Journal of the China Society for Scientific and Technical Information*, 41(06), 549-557

Zhao Z. Y. (2022). Outstanding and Intelligent Empowerment of Scientific and Technological Intelligence in Complex Information Environments. Journal of the China Society for Scientific and Technical Information, 41(12), 1229-1237.

https://doi.org/10.51408/issi2025_107

Exploring Google Books, Open Library, and Wikipedia as Sources for Book Metadata: The UK and Lithuanian Cases

Eleonora Dagienė

eleonora.dagiene@mruni.eu
Institute of Communication, Mykolas Romeris University,
Ateities g. 20, LT-08303 Vilnius (Lithuania)

Abstract

This paper investigates the potential of Google Books, Open Library, and Wikipedia as sources of metadata for scholarly books, focusing on publications from the UK and Lithuania. Utilising ISBNs as unique identifiers, the study analyses the availability, accuracy, and completeness of metadata across these platforms. Initial findings reveal significant disparities between UK and Lithuanian book metadata, with UK publications exhibiting higher coverage and consistency. The research highlights the limitations of these sources, particularly for non-English language publications, and underscores the need for further investigation to develop a more comprehensive and reliable book metadata ecosystem. This research contributes to the ongoing discussion about improving book metrics and enhancing the evaluation of scholarly outputs.

Introduction

Bibliometric research relies heavily on comprehensive and reliable data sources. However, existing research often faces limitations in capturing the full spectrum of scholarly publications, particularly scholarly books, which remain the most challenging and therefore least researched outputs (Borgman & Furner, 2005). Previous studies have explored book citation metrics using various sources, including traditional journal-oriented citation indexes (Halevi et al., 2016; Zuccala & Robinson-García, 2019), online platforms such as Google Books, Google Scholar, and Wikipedia (Kousha et al., 2011; Kousha & Thelwall, 2017), and the WorldCat library catalogue (Torres-Salinas et al., 2021). These studies, however, often focus on books already included in established databases, which may exhibit geographic or linguistic biases.

This limitation becomes evident when examining national research outputs. For instance, a significant proportion of books from countries such as Lithuania (Dagienė, 2024a) and Croatia (Sīle et al., 2021) are entirely missing metadata in the internationally recognised WorldCat catalogue, rendering them practically invisible to international readers. This lack of comprehensive metadata hinders the development of intelligent research metrics (Moed, 2007), and ultimately limits our ability to accurately evaluate and recognise the contributions of scholarly books.

This research-in-progress explores the availability and quality of book metadata in Google Books, Open Library, and Wikipedia for scholarly books submitted as research outputs in the UK and Lithuania, using ISBNs as the primary book identifier. The study aims to answer the following research questions:

1. What is the availability of book metadata in these sources for these particular books (from the UK and Lithuania)?

2. How consistent and accurate is the metadata across these chosen sources?
3. What are the challenges and opportunities in using these data sources for developing book metrics?

By addressing these research questions, this study aims to contribute to a better understanding of the challenges and opportunities in leveraging book metadata for research evaluation and knowledge discovery. Specifically, it explores the possibilities and sources for creating intelligent research metrics. The primary goal is to identify potential data sources and approaches for developing comprehensive intelligent book metrics that can reveal the merit of every book, contributing to a more nuanced, fair, and effective research evaluation system (European Commission, 2021; UNESCO, 2021). Combined with transparent peer review, such intelligent research metrics hold the potential to transform research evaluation practices and address the needs of the future of scholarly communication (Kraker et al., 2016).

Methodology

This research employs a mixed-methods approach, combining quantitative and qualitative analysis of book metadata from three globally recognised sources: *Google Books*[1], *the Open Library*[2], and *Wikipedia*[3]. The empirical analysis uses datasets of books submitted as research outputs for research evaluation between 2008 and 2020, comprising 38,050 ISBNs in the UK (Dagienė, 2023c) and 5,199 ISBNs in the Lithuanian datasets (Dagienė, 2023b). These datasets provided publication years (as provided by submitting institutions). Additionally, they provided book type (authored book or edited volume from the national submission systems), country of ISBN issuance, publisher name, and primary publisher occupation (obtained from the Global Register of Publishers) (Dagienė, 2024b).

Primary metadata (authors, titles, publishers, and publication years) for each ISBN was collected from the three sources using the Python package *isbnlib*[4], which provides functions for retrieving metadata via application processing interfaces (APIs). To track changes in metadata availability, data for all ISBNs from Google Books, Open Library, and Wikipedia were collected in both January 2023 and January 2025.

The analysis focused on determining the number of books from the UK and Lithuanian datasets present in each source and assessing the completeness of their metadata. Books were categorised based on the level of agreement between the three sources. "Matched" shows an exact match in at least two sources, ideally three, suggesting accurate data. "Partial match" signifies potential data availability where at least one author or key title words matched across at least two sources. "One source only" shows data available in only one of the three sources. "No exact match"

[1] Google for Developers. https://developers.google.com/books/docs/v1/libraries accessed 2 January 2025

[2] Open Library. Developer Center / APIs / Books API https://openlibrary.org/dev/docs/api/books accessed 2 January 2025

[3] Wikimedia REST API https://en.wikipedia.org/api/rest_v1/#/Citation accessed 2 January 2025

[4] isbnlib – a python library to validate, clean, transform and get metadata of ISBN strings (for devs) https://github.com/xlcnd/isbnlib; accessed 2 January 2025

highlights discrepancies in metadata elements requiring further review. "No data" means no metadata is available in any of the sources explored.

The figures presented in the following sections illustrate the availability of metadata (author, title, publisher, year, language) for the UK and Lithuanian book ISBNs from 2008 to 2020. The findings suggest that if books have metadata that fall into the first three categories, the sources might be suitable for compiling ISBN metadata for various purposes.

Availability of book metadata

This section presents initial findings on the availability of book metadata in Google Books, Open Library, and Wikipedia. The analysis examines datasets of the UK and Lithuanian book ISBNs, comparing the metadata gathered in January 2023 and January 2025.

Overall, metadata availability for UK books is high across all sources. Google Books' coverage remained consistently high, with 84.3% (32,076 ISBNs) in 2023 and 81.4% (30,972 ISBNs) in 2025. Open Library's coverage increased slightly from 92.1% (35,028 ISBNs) in 2023 to 93.6% (35,601 ISBNs) in 2025. Wikipedia, while initially the highest in 2023 at 99.1% (37,695 ISBNs), saw a decrease to 93.7% (35,655 ISBNs) in 2025, warranting further investigation.

In contrast, Lithuanian book metadata is less readily available. Google Books' coverage increased from 27.6% (1,436 ISBNs) in 2023 to 42.0% (2,181 ISBNs) in 2025, but a significant proportion (58%) still lacks records. Open Library showed minimal change, with coverage around 31-32%. Wikipedia's coverage also decreased, from 70.9% (3,688 ISBNs) in 2023 to 55.4% (2,881 ISBNs) in 2025, possibly because of fewer Lithuanian books being cited/mentioned on Wikipedia.

Combining the 2025 data from all three sources, only 1.5% of UK ISBNs lack metadata. Wikipedia and Open Library are the best choices for UK ISBN metadata, with Google Books also providing sufficient coverage. The picture is less promising for Lithuanian ISBNs, with over a third lacking records even after combining data from all three sources. Even when records exist, they are often incomplete. The next section will explore the consistency and accuracy of the metadata. Further research is needed to understand the decrease in Wikipedia coverage for both UK and Lithuanian books and to identify additional data sources to improve metadata availability for Lithuanian books.

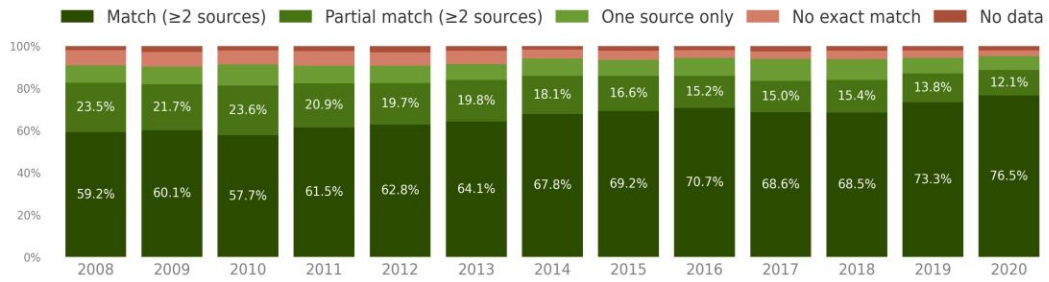
Accuracy and completeness of the metadata

This section examines the accuracy and completeness of key metadata fields (author names, titles, publishers, years, and languages) across Google Books, Open Library, and Wikipedia. Because of space constraints, this paper presents combined results from the three sources. A detailed analysis of each source will be provided in the full paper.

Figure 1 shows how author information is represented, combining data from all three sources. UK ISBNs consistently show higher author information availability than Lithuanian ISBNs, ranging from 90% to 95% for UK books compared to 49% to 57% for Lithuanian books between 2008 and 2020. Furthermore, almost 90% of UK

ISBNs have author information that is consistent in at least two sources in recent years, indicating higher data quality. In contrast, only slightly over 20% of Lithuanian ISBNs have matching author information, with almost half lacking author data across all three sources. This highlights a significant gap in author information for Lithuanian books.

a)



b)

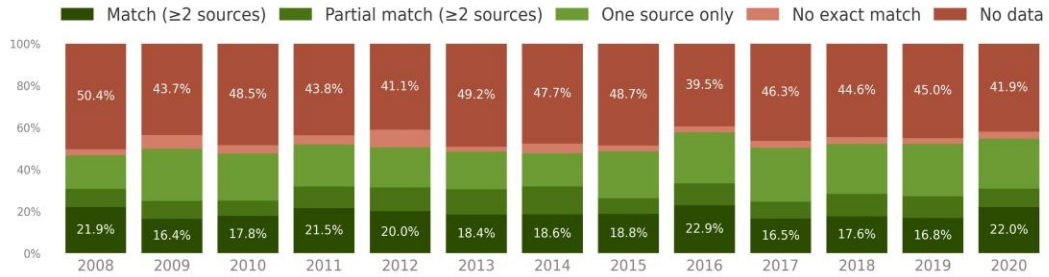
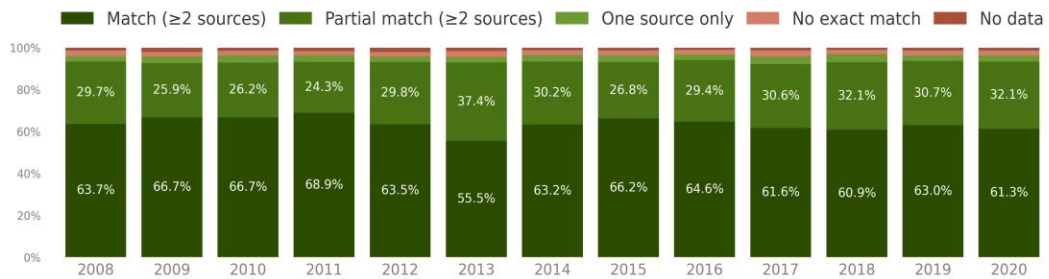


Figure 1. Author information availability for: a) UK and b) Lithuanian ISBNs.

Turning to ***title information***, the sources provide more complete data for titles than for authors (combining both matched and partially matched records). As with author information, UK ISBNs show higher data quality with at least 90% of UK titles have matches in two or three sources. While Lithuanian ISBNs have less missing title data than author data, they still face challenges. Only around 20% of titles are consistently represented in recent years, and over 30% of titles have no data across any of the sources. For Lithuanian ISBNs, information available in only one source is similar to the amount of matched or partially matched records across multiple sources. This suggests potential inconsistencies in title information for Lithuanian books that are already available not speaking about missing records.

a)



b)

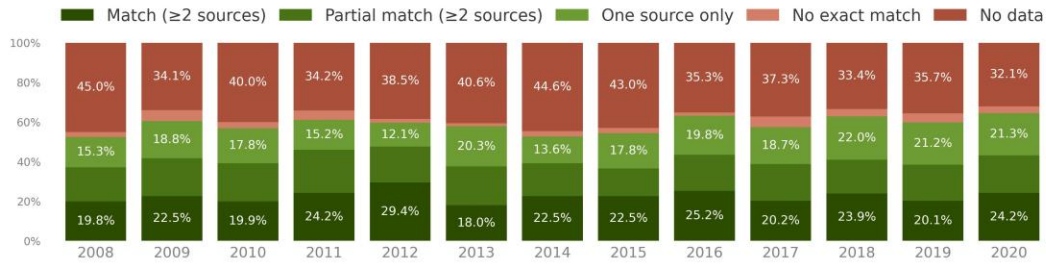
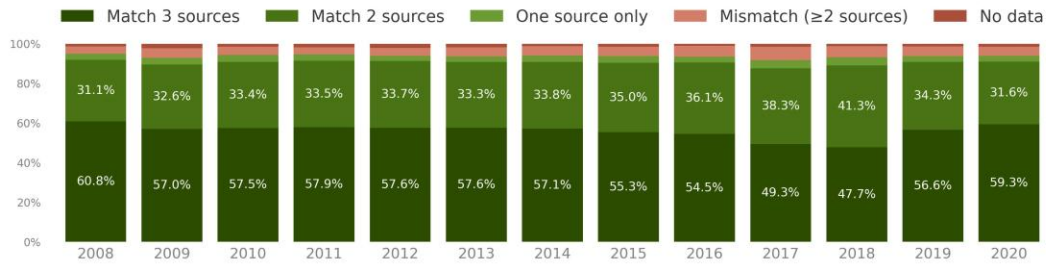


Figure 2. Title information availability for: a) UK and b) Lithuanian ISBNs.

Figure 3 illustrates the availability of year information. Despite more reliable year information from UK ISBNs, we found some inconsistencies in the gathered data; sometimes, the years extracted from the sources were clearly inaccurate (e.g., 2028, 1958, or 1810).

a)



b)

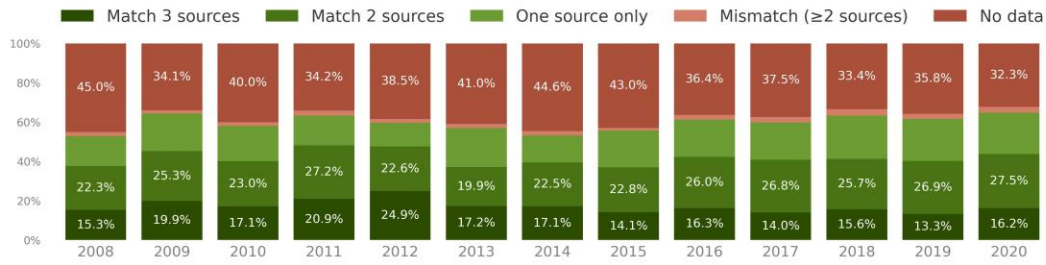


Figure 3. Year information availability for: a) UK and b) Lithuanian ISBNs.

Additionally, some years mismatched between those gathered from three sources and those reported by the UK and Lithuanian institutions at the submission stage. These mismatches were more frequent for UK ISBNs than Lithuanian ISBNs. Despite these issues, over half of the year records for UK ISBNs match across all three sources, and when combined with those that match in two sources, almost 90% of UK records have consistent information on the years. In contrast, over a third of Lithuanian records are missing year information entirely, with very few records having years that match across all three sources.

Publishers. Only 2.1% of UK books lack publisher information in their records, compared to 42.0% of Lithuanian books. Figure 4 shows the overall percentages of

available and missing publisher data, with numbers very similar to those seen for author and title information. A closer look at 'Partial match' records revealed cases where publishers operate under multiple imprints. In such cases, one source may list the imprint's name, while others provide the parent publisher's name, leading to discrepancies. Similar inconsistencies arise after publishers merged, acquired, or reorganised their imprints. Considering that so many ISBNs have no metadata even with all three sources combined, if someone is looking for the reliable publisher data, they can get data from the Global Register of Publishers (Dagienė, 2024b). This registry metadata contains publishers' information for every ISBN issued in any of the over two hundred countries that have joined the ISBN system.

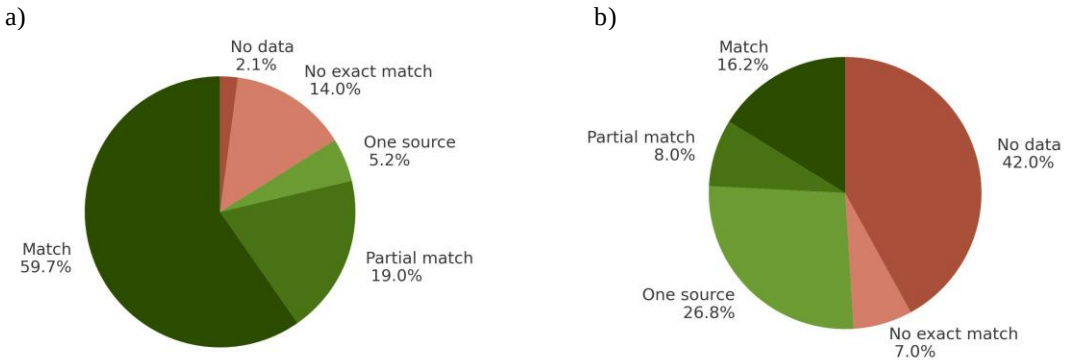


Figure 4. Publisher information availability for: a) UK and b) Lithuanian ISBNs.

Book ***language information*** is available only in Google Books, and its coverage is lower than that of authors, titles, or publishers for both the UK and Lithuanian datasets. The results show that UK books were issued in 29 languages, while Lithuanian books were issued in 17 languages. In the UK ISBN results, language information is missing for 18.6% of books. Of the entire UK dataset, 78.1% are in English, 1.8% are in German, 0.6% are in French, and other languages have even smaller numbers of ISBNs (Figure 5a).

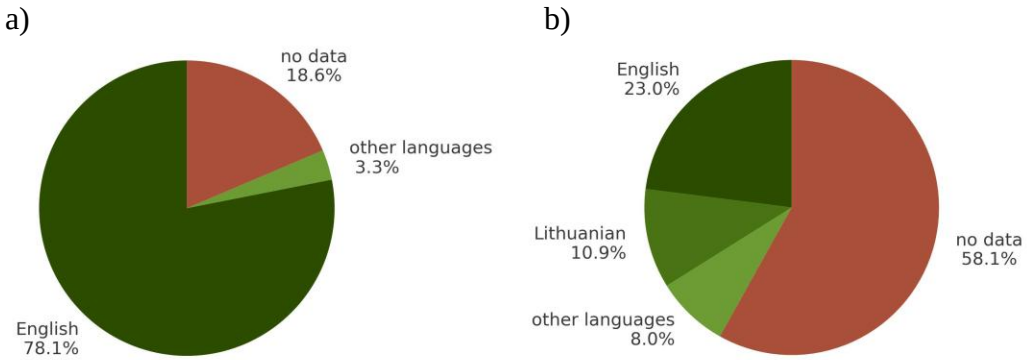


Figure 5. Language information availability in Google Books for: a) UK and b) Lithuanian ISBNs.

Interestingly, Figure 5b shows that more Lithuanian ISBNs are assigned to English-language books than to Lithuanian ones, which contradicts previous findings that if half of the Lithuanian books were issued in Lithuania, they would likely be in Lithuanian, not in other languages. Presumably, the significant of metadata of domestically published books is missing in the international sources analysed in this research study.

This analysis of the accuracy and completeness of metadata across Google Books, Open Library, and Wikipedia reveals significant variations in data quality for UK and Lithuanian books. The UK's ISBNs consistently show higher data quality and completeness across all metadata fields. In contrast, Lithuanian ISBNs exhibit lower data quality, with a significant proportion missing primary metadata elements as author names and titles. These findings highlight the challenges in relying solely on these sources for comprehensive book metadata, particularly for research evaluation purposes. The full paper will provide a more detailed analysis of each source and explore strategies to address these limitations.

Challenges, opportunities, and initial conclusions of using three data sources

This research investigates the potential of Google Books, Open Library, and Wikipedia as sources for book metadata, empirically testing their efficacy using UK and Lithuanian scholarly publications. While initial results show these platforms providing a significant amount of ISBN metadata, they also reveal notable limitations, particularly for publications from non-English-speaking countries as Lithuania. This disparity is clear in the higher coverage and accuracy of metadata for UK books compared to their Lithuanian counterparts. Moreover, inconsistencies in author names, publication years, and publisher information frequently occur across these sources, even for already represented books. These initial findings underscore the need for further investigation to identify the missing data and find out the reasons behind these gaps.

The full paper will, therefore, focus on three key areas. First, a deeper analysis will be conducted to investigate the influence of publisher type and size on metadata availability and quality, exploring potential avenues for improving the representation of books and their ISBNs through publisher engagement. Second, the research will identify the underlying sources that Google Books, Open Library, and Wikipedia leverage to compile their extensive book databases and potentials for covering underrepresented book metadata. Third, a thorough examination of the nature and the extent of discrepancies across these platforms will be undertaken. This will provide crucial insights into whether these inconsistencies can be rectified, paving the way for a unified and more reliable book metadata universe.

Ultimately, the overarching goal of this project is to identify and analyse a comprehensive range of book metadata within three platforms explored here, representing only the initial steps. By researching the ways for integrating data from the diverse sources, we aim to achieve maximum metadata coverage for books within any dataset, moving beyond the current limitations of a Western, English-language-centric focus and embracing a truly global perspective. This will significantly

enhance the reliability and comprehensiveness of book metadata, improving its value for research evaluation (Dagienė, 2023a) and the development of robust book metrics that assist book peer-review assessment.

References

Borgman, C. L., & Furner, J. (2005). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*(1), 2–72. https://doi.org/10.1002/aris.1440360102

Dagienė, E. (2023a). Prestige of scholarly book publishers—An investigation into criteria, processes, and practices across countries. *Research Evaluation*, *32*(2), 356–370. https://doi.org/10.1093/reseval/rvac044

Dagienė, E. (2023b). *The metadata of books submitted as research outputs to annual Lithuanian research assessments from 2008 to 2020 [Data set]* [Dataset]. https://doi.org/10.5281/zenodo.10070933

Dagienė, E. (2023c). *The metadata of books submitted as research outputs to REF 2014 and REF 2021 [dataset]* [Csv]. Zenodo. https://doi.org/10.5281/zenodo.10071003

Dagienė, E. (2024a). Mapping scholarly books: Library metadata and research assessment. *Scientometrics*, *129*, 5689–5714. https://doi.org/10.1007/s11192-024-05120-1

Dagienė, E. (2024b). The challenge of assessing academic books: The UK and Lithuanian cases through the ISBN lens. *Quantitative Science Studies*, *5*(1), 98–127. https://doi.org/10.1162/qss_a_00284

European Commission. (2021). *Towards a reform of the research assessment system* (p. 21). European Union. https://doi.org/10.2777/707440

Halevi, G., Nicolas, B., & Bar-Ilan, J. (2016). The complexity of measuring the impact of books. *Publishing Research Quarterly*, *32*(3), 187–200. https://doi.org/10.1007/s12109-016-9464-5

Kousha, K., & Thelwall, M. (2017). Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, *68*(3), 762–779. https://doi.org/10.1002/asi.23694

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147–2164. https://doi.org/10.1002/asi.21608

Kraker, P., Dörler, D., Ferus, A., Gutounig, R., Heigl, F., Kaier, C., Rieck, K., Šimukovič, E., & Vignoli, M. (2016, December 30). *The Vienna Principles: A Vision for Scholarly Communication in the 21st Century*. https://doi.org/10.5281/zenodo.55597

Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, *34*(8), 575–583. https://doi.org/10.3152/030234207X255179

Sīle, L., Guns, R., Zuccala, A., & Engels, T. (2021). Towards complexity-sensitive book metrics for scholarly monographs in national databases for research output. *Journal of Documentation*, *77*(5), 1173–1195. https://doi.org/10.1108/JD-06-2020-0107

Torres-Salinas, D., Arroyo-Machado, W., & Thelwall, M. (2021). Exploring WorldCat identities as an altmetric information source: A library catalog analysis experiment in the field of Scientometrics. *Scientometrics*, *126*(2), 1725–1743.

https://doi.org/10.1007/s11192-020-03814-w

UNESCO. (2021). *Recommendation on Open Science* (p. 34). UNESCO.

Zuccala, A., & Robinson-García, N. (2019). Reviewing, Indicating, and Counting Books for Modern Research Evaluation Systems. In *Springer Handbook of Science and Technology Indicators* (pp. 715–728). https://doi.org/10.1007/978-3-030-02511-3_27

https://doi.org/10.51408/issi2025_141

Exploring Research Collaboration of Private Universities in Emerging EU Countries: A Comparison with Public Sector

Alexander Dmitrienko[1], Nataliya Matveeva[2], Maria Yudkevich[3]

[1]*a.dmitrienko@e-kvadrat.com*, [2]*n.matveeva@e-kvadrat.com*
E-Quadrat Science & Education (Germany)

[3]*myudkevic@univ.haifa.ac.il*
University of Haifa (Israel)

Abstract

We study the research performance of research-active universities in the five countries with European Union candidate status (Albania, Bosnia and Herzegovina, Georgia, North Macedonia, and Serbia). We examine to what extent "research active" private HEIs differ in their research activities from public ones and why. Using knowledge capabilities theory, we demonstrate how patterns of national and international co-authorship can explain the survival strategies of private universities and their position in the academic markets of these countries today. Based on the publication data 2010-2022 from Scopus, we analyze the characteristics of universities' publication output and their scientific collaboration. The Wilcoxon signed-rank test was applied to assess significant differences between the two university groups. To estimate the similarity of universities' collaboration patterns, we apply the blockmodeling procedure to both non-normalized and normalized (using Balassa normalization) co-authorship networks of the universities. We reveal that private and public universities demonstrate similar characteristics in publication output and scientific collaboration. They are statistically different only in size, measured by the number of students and scientific staff, value of publication output, and the number of papers produced independently. Private universities almost do not collaborate with each other inside the country; their collaboration is skewed towards one or two public universities. Moreover, the position of private universities within the national academic network is often peripheral, and they do not fully realize their potential for collaboration. Our study reveals that private universities in the analyzed countries tend to mimic existing public ones in their research activities, adopting similar research practices.

Introduction

Private sectors of higher education systems in the last several decades have experienced a rapid growth (Levy, 2018). In different countries, they vary in size and functions (Reisz, Stock, 2012) and therefore relate to the public sector in various ways (e.g. complementing it in empty niches, competing with - Teixeira et al., 2012). Accordingly, depending on the relationships between the private and public sectors and their relative roles within the national systems, higher education institutions (HEI) from different sectors may constitute a homogenous group or perform differently within one country (Teixeira et al., 2017).

In most countries, the private sector has evolved evolutionarily and has a relatively long history (Levy, 2024; Altbach, Levy, 2005). However, in some countries and regions, it has emerged relatively recently due to significant changes in legislation and the changing political-economic context (Spain or the region of interest - Brankovic, 2014; Casani et al., 2014). It is not uncommon to see that low dynamics in the number of public institutions is accompanied by an explosion and consequent

decline in the number of private organizations. Indeed, for various reasons, many of them do not survive (for the case of Russian universities, see Kuzminov, Yudkevich, 2022).

What is the place of the private HEI sector in the national system after such an initial "rapid growth phase", and to what equilibrium state does the system converge? What differences do we see between "surviving" private HEIs and public HEIs? Existing studies predominantly focus on the teaching aspects of these differences. In our paper, we answer these questions by focusing on the research component of university performance.

We use network capabilities theory (Eisenhardt, Martin, 2000; Ritter et al., 2002) to explain the patterns of university collaborations as instruments to acquire additional knowledge capital and general embeddedness in the academic market. While in existing literature this theory is mainly applied to business firms (Mitrega et al., 2011; Sullivan, Weerawardena, 2006) and only few papers make an attempt to use it for an analysis of university strategies (King'oo et al., 2020; Huang, 2014), we demonstrate that this theory can be a powerful tool for analysis of organizations centered around human capital.

Data and Methodology

We study universities with a non-zero research output in international journals indexed in the Scopus database. We selected a relatively long time period of 12 years, 2010-2022, to form robust publication statistics for the analysis. Our sample consists of 43 public and private universities from five EU enlargement countries that by 2023 had candidate status to join the European Union: Albania, Bosnia and Herzegovina, Georgia, North Macedonia, and Serbia. In the sample, we include all universities with at least one publication in the analyzed period. Of the 43 universities studied, 14 are private, and 29 are public. For these universities, we collected bibliometric data on publications (articles and reviews) related to their profiles in the Scopus database, as well as data from open internet sources (QS World University Rankings, WHED by IAU, Rankless by CCL, universities' official websites, etc.).

We use variables that characterize overall and per capita publication output to analyze publication activity. For scientific collaborations, we include variables that characterize collaboration at the author, organizational, and country levels. All variables are presented in Table 1. To determine whether the observed differences between public and private university groups are statistically significant, we apply the Wilcoxon signed-rank test (Woolson, 2005). In the final stage of our analysis, we assess both the collaborative proximity and the similarity of collaboration patterns between public and private universities across countries. Collaborative proximity is measured by the share of joint publication output and key research fields, while similarity is evaluated using an indirect blockmodeling procedure applied to co-authorship networks (where nodes are universities and links represent joint publications). Given the substantial differences in publication volume between universities, we apply Balassa normalization. Compared with other normalization methods (e.g., Jaccard and Affinity normalization), Balassa normalization is less

sensitive to unit size. It allows for the estimation of the collaboration potential of the analyzed units (Matveeva, Batagelj, and Ferligoj 2023). All computations were done using the programs R (R Core Team (2023)) and Pajek (http://mrvar.fdv.uni-lj.si/pajek/).

Results

We observe that the essential characteristics of public and private universities are similar. Figure 1 shows the analyzed characteristics' median value and distribution within each university group. Private universities have slightly fewer publications per person than public ones, with a median value of 2 papers per person for private universities compared to 2.7 for public universities. Even though private universities more often have fewer publications than public universities, they demonstrate a relatively high share of high-quality output. The median share of Q1 publications in private universities is 32%, while in public universities, it is 30%, which indicates that private universities also focus on high-quality research. Another similar characteristic of the two university groups is the number of students per scientific staff: both groups have a median value of about 15 students per person. This means that the teaching load of staff is quite similar in public and private universities. At the same time, private universities have a higher share of dominant research fields compared to public universities, which describes them as more specialized universities. For private universities, the median value of the share of the dominant research area is 20%, while for public universities, it is about 5%. At the end of the next section, we estimate the significance of the observed differences.

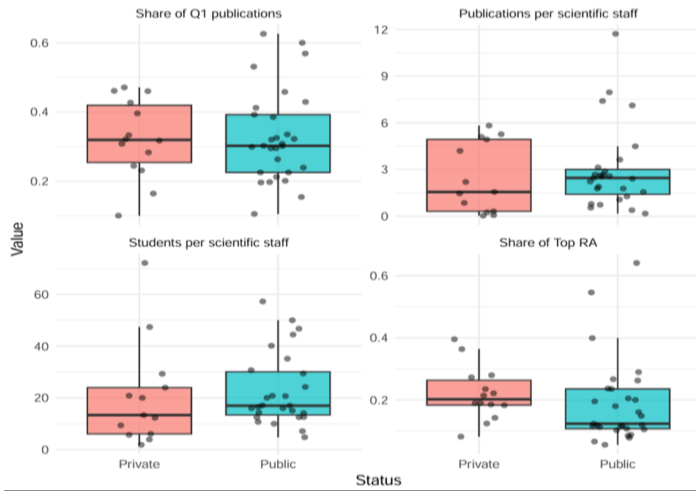


Figure 1. Publication characteristics of public and private universities. The line inside the boxes represents the median value, the size of the boxes covers 50% of the observations.

We find that at all three analyzed levels public and private universities are very similar in the share of papers prepared in co-authorship (Fig.2). Only the share of papers prepared by the university itself is a bit higher in the public university sector:

38% when in the private group it is 29%. Both university groups have a share of single-authored works, about 10%. The share of national collaboration is about 30% in both groups. Thereby, we observe that public and private universities actively collaborate with other countries and other organizations (about 70% of publications), and most papers (90%) are prepared in co-authorship.

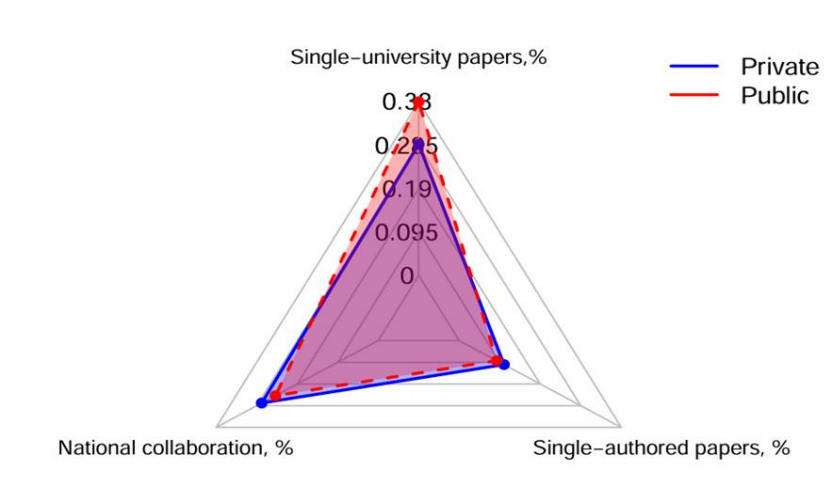


Figure 2. Share of publications without collaboration on different levels (author, organization, country).

We apply the Wilcoxon test to answer whether the observed difference between public and private universities is statistically significant. The results of the Wilcoxon test demonstrate that public and private universities in our sample differ statistically in size: the number of publications, the number of scientific staff, and the number of students (Table 2). Among the collaboration characteristics, only the share of publications with a single affiliation is statistically significant: public universities have more publications prepared independently. This result is quite logical: public universities are bigger than private ones, so they have the capacity to produce research publications on their own. Other characteristics become non-significant, which means that their variation cannot be explained by university status.

Table 1. Wilcoxon signed-rank test results.

Variable	Private_vs_Public universities
Number of publications	0.0016*
Number of scientific staff	0.0001*
Number of students	0.0001*
Students per person	0.2788
Share of Q1 publications	0.7361
International collaboration. %	0.8256
Average number of authors per work	0.0634
Number of publications per scientific staff	0.4238
Share of publications with 1 affiliation	0.0292*
Share of top research area	0.0848
Share of publications with 1 author	0.3782
Average number of affiliations in the work with 1 author	0.0900

*Significant at 0.05 level

In Fig. 3, the universities are grouped into clusters according to the similarity of their collaboration patterns with others in the network. We observe that a clear core-periphery structure is evident in Georgia, Serbia, and North Macedonia. There is a core of well-collaborating universities and a periphery of universities that collaborate very weakly with others. Private universities in these countries are typically located in the periphery. In Albania, and Bosnia and Herzegovina, a different collaboration structure is observed: there is a core, a semi-core, and a periphery. The core consists of universities that collaborate extensively, the semi-core includes universities that actively collaborate with both the core and the periphery, and the periphery comprises universities that collaborate weakly. Private universities in our sample are in the same clusters, which indicates the similarity of their collaboration patterns. However, their cluster position is often in the periphery. Only two private universities in Bosnia and Herzegovina are located in the semi-core.

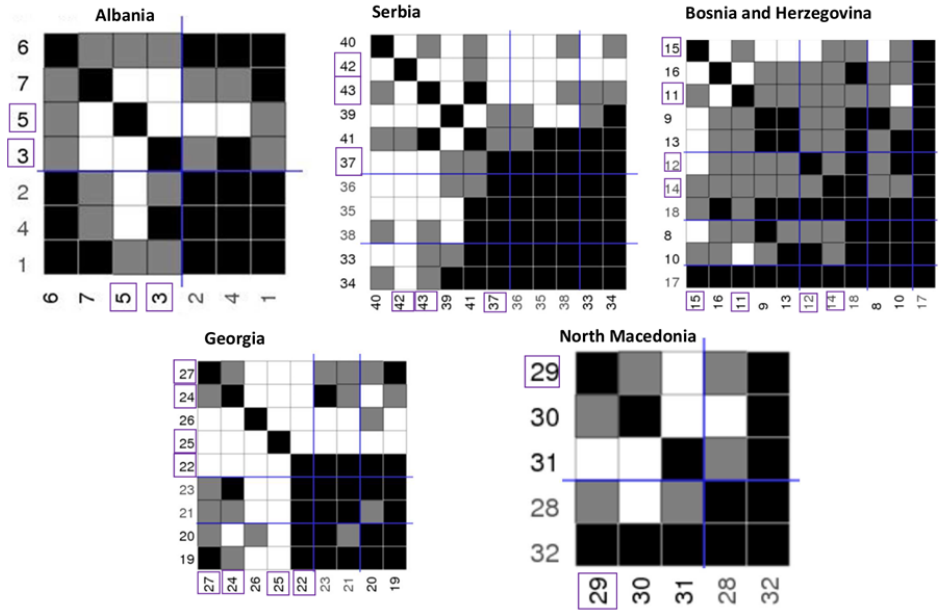


Figure 3. Blockmodeling of universities' collaboration inside the countries. Private universities are in purple frame. White cells represent no publication, grey from 1 to 10, black 10 and more.

Discussion and Conclusion

Our analysis demonstrates that research-active public and private universities in countries under consideration differ significantly in size, measured by a number of scientific staff and students. The lower number of publications in private universities is explained by their size: private universities are smaller. Individual research productivity (measured by the number of publications per person) and the share of most impactful publications (measured by the share of Q1 publications) are similar in both groups of universities. Private and public universities also have similar collaboration characteristics measured via the share of publications prepared in co-authorship on individual, organizational, and national levels. The research competitiveness of private universities can be attributed to their capacity to adapt and strategically reallocate resources and build effective research governance structures, aligning closely with the dynamic capabilities theory.

We observe that private universities primarily collaborate with public ones and almost do not collaborate at all with other private universities. With that, the 'follow by leaders' strategy results in private universities taking a peripheral position in the country's academic network, with reduced independent access to resources. With all its disadvantages, a peripheral position still allows private universities to adopt the experience of public universities and build a research background for future development. A peripheral position provides limited access to material and symbolic

resources (Fumasoli, Barbato, and Turri 2020); as a result, private universities have a reduced capacity to support their research activities.

Our study is limited to research-active universities; hence, we do not include institutions without non-occasional Scopus publications. However, Scopus does not comprehensively cover local national journals, particularly those published in national languages. Moreover, many private universities have very few publications. Consequently, some observed characteristics are not statistically significant and may be associated with more profound underlying differences. These limitations should be addressed in future research.

We conclude that private universities in analyzed countries enroll in the research system by mimicking public universities rather than filling empty niches. Such a mimicry strategy is also observed in other countries with developing academic sectors, for instance, the UAE (Ashour & Kleimann, 2024). Research activities allow private universities to gain legitimacy and elevate their status, and they actively use collaborations as a resource for development. We contribute to the literature by explaining the survival strategies of private universities in countries with relatively new private sectors. Our findings will allow for the design of evidence-based policy measures and initiatives aimed to support collaborative inter-institutional research and to provide an impact toward the balanced development of higher education national systems in a broader European context.

References

Altbach, P. G., & Levy, D. C. (2005). Private higher education: A global revolution (Vol. 2). Brill.

Ashour, S., & Kleimann, B. (2024). Private higher education: a comparative study of Germany and the United Arab Emirates. Research Papers in Education, 39(4), 668-684.

Brankovic, J. (2014). Positioning of private higher education institutions in the Western Balkans: emulation, differentiation and legitimacy building. In The re-institutionalization of higher education in the Western Balkans: the interplay between European ideas, domestic policies, and institutional practices (Vol. 5, pp. 121-144). Peter Lang.

Casani, F., De Filippo, D., García-Zorita, C., & Sanz-Casado, E. (2014). Public versus private universities: Assessment of research performance; case study of the Spanish university system. Research evaluation, 23(1), 48-61.

Eisenhardt, K. M., & Martin, J. A. (2000). Dynamic capabilities: what are they?. Strategic management journal, 21(10-11), 1105-1121.

Fumasoli, T., Barbato, G., & Turri, M. (2020). The determinants of university strategic positioning: a reappraisal of the organisation. Higher Education, 80(2), 305-334

Huang, J. S. (2014). Building Research Collaboration Networks--An Interpersonal Perspective for Research Capacity Building. Journal of Research Administration, 45(2), 89-112.

King'oo, R. N., Kimencu, L., & Kinyua, G. (2020). The role of networking capability on organization performance: A perspective of private universities in Kenya. Journal of Business and Economic Development, 5(3), 178-186.

Kuzminov, Y., & Yudkevich, M. (2022). Higher education in Russia. John Hopkins University Press Press.

Levy, D. C. (2018). Global private higher education: an empirical profile of its size and geographical shape. Higher Education, 76, 701-715.

Levy, D. C. (2024). A world of private higher education. Oxford University Press.
Matveeva, N., Batagelj, V., & Ferligoj, A. (2023). Scientific collaboration of post-Soviet countries: the effects of different network normalizations. Scientometrics, 128(8), 4219-4242.
Mitrega, M., Ramos, C., Forkmann, S., & Henneberg, S. C. (2011). Networking capability, networking outcomes, and company performance. In Proceedings of the IMP Conference.
Reisz, R. D., & Stock, M. (2012). Private higher education and economic development. European Journal of Education, 47(2), 198-212.
Ritter, T., Wilkinson, I. F., & Johnston, W. J. (2002). Measuring network competence: some international evidence. Journal of Business & Industrial Marketing, 17(2/3), 119-138.
Sullivan, G., & Weerawardena, J. (2006). Networking capability and international entrepreneurship: How networks function in Australian born global firms. International marketing review, 23(5), 549-572.
Teixeira, P., Kim, S., Landoni, P., & Gilani, Z. (2017). Rethinking the public-private mix in higher education: Global trends and national policy challenges. Springer.
Teixeira, P., Rocha, V., Biscaia, R., & Cardoso, M. F. (2012). Myths, beliefs and realities: Public-private competition and program diversification in higher education. Journal of Economic Issues, 46(3), 683-704.
Woolson. R. F. (2005). Wilcoxon signed-rank test. Encyclopedia of Biostatistics. 8.

https://doi.org/10.51408/issi2025_103

Exploring the Policy Impact and Funding Mechanisms of Scientific Collaboration Between Taiwan and New Southbound Policy (NSP) Priority Countries

Pei-Ying Chen[1], Tzu-Kun Hisao[2], Cassidy R. Sugimoto[3]

[1]*peiychen@iu.edu*
Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington (USA)

[2]*tkhsiao2@illinois.edu*
School of Information Sciences, University of Illinois Urbana-Champaign (USA)

[3]*sugimoto@gatech.edu*
Jimmy and Rosalynn Carter School of Public Policy, Georgia Institute of Technology (USA)

Abstract

International scientific collaboration defines the fourth age of research, with policy incentives frequently cited as key motivators for researchers to engage in cross-border collaboration and exchange. However, empirical evidence from non-Western contexts remains limited, and the heterogeneity within international collaboration is often overlooked. To address these empirical and conceptual gaps, this study examines the impact of Taiwan's New Southbound Policy (NSP) on its scientific collaboration with eight designated priority countries over the period 2011–2021. Drawing on bibliographic data from the Web of Science (WoS) Extended API, we analyzed 28,465 co-authored articles. Funding status was identified through funding acknowledgments, and co-authorship types were categorized based on the country affiliations of first and last authors. Our preliminary findings show no strong evidence that the NSP itself contributed to the post-NSP growth of scientific collaboration between Taiwan and NSP priority countries. However, we observe a decline in minimal collaborations and an increase in co-affiliated ones, with the former particularly evident in the number of co-publications funded by Taiwan. Despite the null results, this work contributes to the literature by empirically evaluating the effectiveness of science diplomacy initiatives and pointing to their potential limitations.

Introduction

International scientific collaboration (ISC) defines the fourth age of research (Adams, 2013), with policy incentives often cited as key motivators for researchers to engage in cross-border collaboration and exchange (Katz & Martin, 1997). For instance, in response to grand challenges that transcend national borders and the shifting international order, ISC has also gained traction amid renewed interest in science diplomacy (Royal Society, 2010). However, empirical evidence remains scarce and predominantly in the EU context (e.g., Glänzel et al., 1999; Makkonen & Mitze, 2016). Moreover, the heterogeneity within international collaboration has only recently gained attention, particularly with the rise of multiple institutional affiliations (Hottenrott et al., 2021) and the prevalence of shared heritage collaboration (Gök & Karaulova, 2023) as inferred from author surnames (Karaulova et al., 2019). To address these empirical and conceptual gaps, this study examines scientific collaboration between Taiwan and designated priority countries—

including Indonesia, Malaysia, the Philippines, Thailand, Vietnam, Singapore, India, and Australia—under the New Southbound Policy (NSP), considering variations across co-authorship types.

Launched in 2016 as Taiwan's new "Regional Strategy for Asia", the NSP aims to strengthen ties with Indo-Pacific countries amid shifting global and regional geopolitics (Office of the President Republic of China (Taiwan), 2017). The same year, the National Science Council established the Southbound Science & Technology Cooperation (NSTC) project office to (1) promote regional academic cooperation, (2) promote talent exchange and cultivation, (3) build international collaboration platforms, and (4) connect international science parks. This framework provides a unique opportunity to explore the policy's impact on scientific collaboration in a non-Western context.

Specifically, we pose two research questions: (1) How has the New Southbound Policy (NSP) influenced the volume of co-publications between Taiwan and the NSP priority countries? and (2) What funding mechanisms support international scientific collaboration under the NSP, given that research grants are one of the most common R&D policy instruments (Martin, 2016)? Recognizing the complex dynamics involved in collaboration—which reflect not only S&T capacity but also hierarchies within global science (Miao et al., 2024)—we further examine whether the policy's impact and funding mechanisms vary across types of co-authorship.

Data and Methods

For this study, we draw bibliographic data from the Web of Science (WoS) Extended API to retrieve publications published between 2011 and 2021 and having at least one author affiliated with Taiwan and one with at least one of the NSP priority countries. This timeframe provides a five-year window before and after the launch of the NSP in 2016, allowing us to examine changes in collaboration trends and patterns. After cleaning and processing, the analytic sample consists of 28,465 articles written in English, without missing country affiliations for the first and last author, and authorship order not in alphabetical sequence for four or more authors.

The co-authorship types were determined by the country affiliations of first and last authors into *TWN-led* (either first or last authors are affiliated with Taiwan but not NSP priority countries), *NSP-led* (opposite of TWN-led), *Equal* (either first author is affiliated with TWN and last author with NSP or vice versa), *Minimal* (neither first nor last authors are affiliated with TWN or NSP), and *Co-affiliated* (either first or last authors are affiliated with both TWN and NSP).

Funding status was identified through funding acknowledgements. To determine whether an article was funded by Taiwan, we first used Stanza, a natural language processing (NLP) toolkit developed by the Stanford NLP group (Qi et al., 2020), to identify named entities and their types from the funding text. For this work in progress, we focused on geopolitical entities (GPE), such as countries, cities, or states. We then developed a rule-based system that leveraged our knowledge of various ways Taiwan might be referenced in the funding text (e.g., "R.O.C") and included names of Taiwan's cities sourced from the Simplemaps' World Cities database (Simplemaps, n.d.). Articles having at least one named entity pointing to

Taiwan were classified as Taiwan-funded. Finally, we applied regular expressions to cross-check and maximize the number of identified articles.

We fitted piecewise linear regression models to examine the policy effect and funding mechanisms in terms of both absolute and relative changes. Changes before and after the launch of the NSP was described by two separate slopes. We also included interaction terms between co-authorship type and each slope to assess whether the changes varied by co-authorship type. To test whether the pre- and post-NSP slopes differed significantly, the pre-NSP slope was re-specified as a linear time slope covering the entire period from 2011 to 2021. To facilitate interpretation, the time variable was centered at 2016, and TWN-led was used as the reference category, so the intercept represents the expected number/share of TWN-led co-publications in 2016. Absolute changes in co-publication counts were modeled using a negative binomial distribution to account for skewness. Given the nested structure of the data, which introduces dependence among observations, we employed robust standard errors to account for clustering.

Preliminary Results

Policy Impact

As shown in Table 1, TWN-led co-publication counts (the reference group) increased by 12% annually prior to 2016 ($p < .001$) and by 19% annually after the implementation of NSP in 2016 ($p < .001$) (M1). The 6% difference between the pre- and post-NSP periods, however, is not statistically significant (model not shown). As of 2016, minimal collaboration occurred 1.16 times as frequently as TWN-led ones ($p = .034$), while co-affiliated co-publications were only half as frequent ($p < .001$). Although the pre-NSP growth rates did not differ by co-authorship type, post-NSP, minimal collaboration exhibited a slower growth trajectory, with an additional annual decrease of 5% ($p = .097$) while co-affiliated co-publications grew more rapidly, with an additional 10% annual increase ($p = .047$).

Table 1. Results from piecewise linear models for co-publications and co-publications funded by Taiwan.

	Co-publications		Taiwan-funded co-publications	
	M1 (N)	M2 (%)	M3 (N)	M4 (%)
(Intercept)	453.98***	21.62***	268.44***	59.09***
	(21.57)	(0.93)	(17.87)	(3.29)
Pre-NSP	1.12***	0.02	1.16***	1.77
	(0.02)	(0.30)	(0.03)	(1.07)
Post-NSP	1.19***	-0.23	1.20***	0.88
	(0.03)	(0.32)	(0.04)	(0.92)
Co-authorship				
TWN-led	–	–	–	–
NSP-led	0.95	-1.17	0.31***	-40.13***

	(0.08)	(1.27)	(0.05)	(4.07)
Equal	1.06	1.23	0.69**	-21.23***
	(0.07)	(1.38)	(0.09)	(4.41)
Minimal	1.16*	2.99+	1.03	-7.85+
	(0.08)	(1.49)	(0.10)	(4.17)
Co-affiliated	0.49***	-11.16***	0.36***	-16.16**
	(0.06)	(1.24)	(0.04)	(5.34)
Pre-NSP × Co-authorship				
Pre-NSP × NSP-led	1.01	0.15	1.00	-1.56
	(0.03)	(0.55)	(0.05)	(1.23)
Pre-NSP × Equal	0.98	-0.46	0.97	-1.05
	(0.02)	(0.50)	(0.05)	(1.77)
Pre-NSP × Minimal	1.00	-0.17	0.94	-3.45*
	(0.04)	(0.98)	(0.05)	(1.37)
Pre-NSP × Co-affiliated	1.05	0.37	0.99	-2.91+
	(0.04)	(0.38)	(0.04)	(1.70)
Post-NSP × Co-authorship				
Post-NSP × NSP-led	1.06	1.29**	0.98	-1.85
	(0.04)	(0.44)	(0.05)	(1.11)
Post-NSP × Equal	0.98	-0.48	0.96	-0.68
	(0.03)	(0.42)	(0.04)	(1.20)
Post-NSP × Minimal	0.95+	-1.01+	0.83***	-5.57***
	(0.03)	(0.56)	(0.03)	(1.16)
Post-NSP × Co-affiliated	1.10*	1.33**	1.09*	-0.25
	(0.05)	(0.41)	(0.04)	(2.20)
Num.Obs.	55	55	55	55
R2 / R2 Adj.		0.924 / 0.897		0.944 / 0.924
RMSE	42.82	1.37	22.79	3.54

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Results from modeling percentage changes (M2) largely aligns with the observations above, though the pre-/post-NSP slope changes are not significant, suggesting that the shares of TWN-led co-publications remain relatively stable. In addition to the varying post-NSP growth patterns observed in minimal (-1.01%, $p = .078$) and co-affiliated (1.33%, $p = .002$) co-publications, there also appears to be an additional increase in the share of NSP-led ones, which grew by 1.29% annually more than that of TWN-led ones ($p = .006$).

It should be noted, however, that the changes pertain only to the models with two direct slopes, each compared independently against the intercept. No significant

differences were found between the pre-NSP and post-NSP periods. The predicted trends of co-publications between Taiwan and NSP priority countries, in terms of both absolute counts and relative shares, are shown in Figure 1.

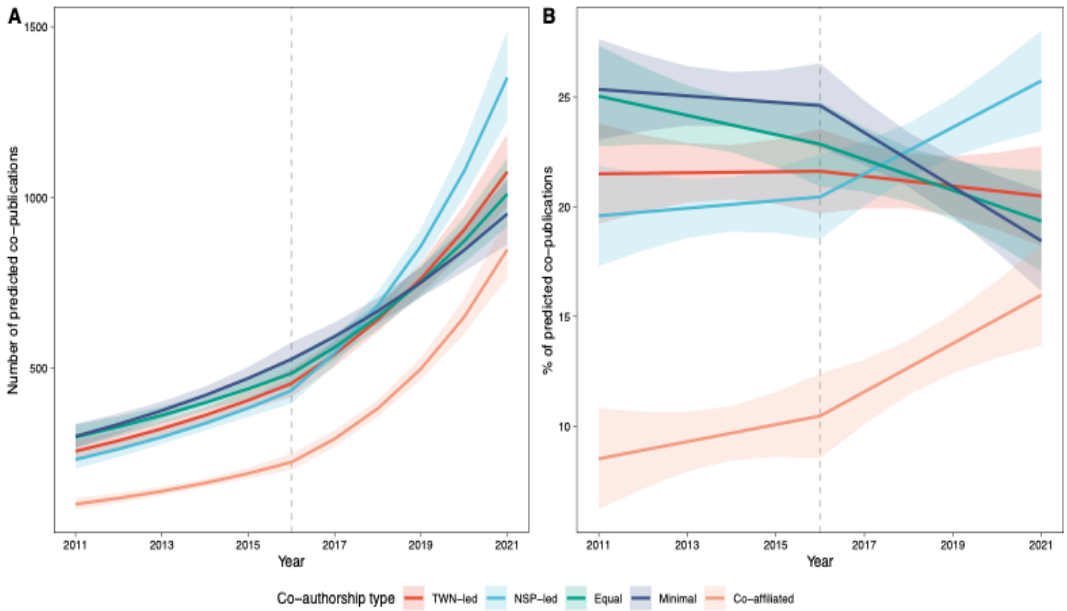


Figure 1. Predicted trends in co-publications between Taiwan and NSP priority countries (2011–2021), by co-authorship type: (A) Absolute changes in counts; (B) Relative changes in shares. The grey dashed line marks the year 2016, when the NSP was launched.

Funding Mechanisms

For TWN-led co-publications funded by Taiwan, the annual growth rates were 16% before and 20% after the NSP launched in 2016 (M3), although the 4% difference is not statistically significant (model not shown). All other co-authorship types, except for minimal collaboration, had significantly fewer papers compared to TWN-led co-publications as of 2016. Similar to M1, differences in growth rates among co-authorship types became more pronounced post-NSP, with minimal collaboration being 17% more slowly ($p < .001$), while co-affiliated publications grew 9% more quickly ($p = .037$). It is worth noting that the pre-/post-NSP difference in minimal collaboration is statistically significant at 0.01 level (IRR = .87, $p = .082$).

When looking at the proportion of TWN-fund co-publications among all funded papers (M4), the share of TWN-led papers funded by Taiwan increased only slightly, by 1.77% annually pre-NSP and 0.88% post-NSP, and neither is statistically significant. In 2016, all other co-authorship types had substantially lower shares: NSP-led, equal, co-affiliated, and minimal collaborations were 40.13% ($p < .001$), 21.23% ($p < .001$), 16.16% ($p = .004$), and 7.85% ($p = .067$) lower, respectively, compared to TWN-led co-publications. Even during the pre-NSP period, the shares of minimal and co-affiliated co-publications saw additional annual declines of 3.45% and 2.91%, respectively. Post-NSP, the share of minimal collaboration dropped by

an additional 5.57% relative to TWN-led co-publications ($p < .001$), while the additional loss for co-affiliated ones was much smaller, at just 0.25%.

The predicted trends of TWN-funded co-publications between Taiwan and NSP priority countries, in terms of both absolute counts and relative shares, are presented in Figure 2. Particularly notable is the flattening of minimal collaboration funded by Taiwan after 2016 accompanied by a sharp decline in its shares. Also noteworthy is the share of TWN-funded co-affiliated publications, which, like minimal collaborations, showed similar downward trend prior to 2016, but experienced growth comparable to that of TWN-led co-publications following the launch of the NSP.

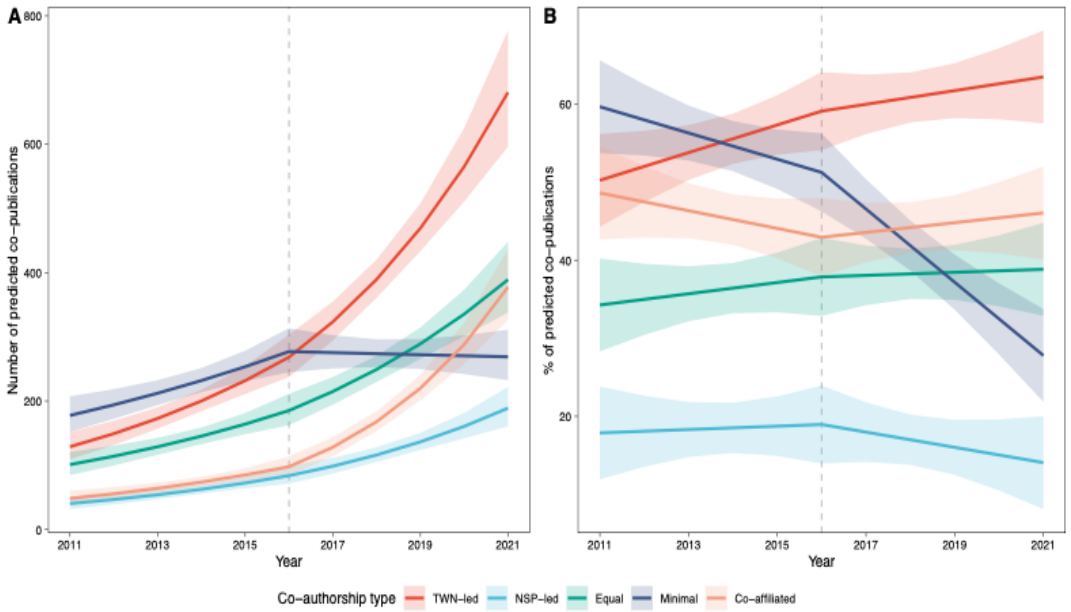


Figure 2. Predicted trends in Taiwan-funded co-publications between Taiwan and NSP priority countries (2011–2021), by co-authorship type: (A) Absolute changes in counts; (B) Relative changes in shares. The grey dashed line marks the year 2016, when the NSP was launched.

Discussion and Tentative Conclusion

Statistically, we found no strong evidence that the NSP itself contributed to the overall growth of scientific collaboration, as measured by co-publications between Taiwan and NSP priority countries. However, we did observe variations across co-authorship types: minimal collaboration and co-affiliated publications displayed distinct post-NSP patterns, with the former declining and the latter increasing. This is especially evident in the number of minimal collaborations funded by Taiwan.

We acknowledge the potential misclassification of Taiwan-based funding using the rule-based approach, which may have introduced bias into the modeling results presented here. In future work, we plan to incorporate metadata for research organizations from Research Organization Registry (ROR), including location (country) and name variants. By leveraging the similarly between embedding

representations of organization names in the WOS and ROR data, we aim to improve the accuracy of identifying the countries affiliated with funding agencies, under the assumption that name variants of an institution will be located near each other in the embedding space.

We also recognize the substantial variation in science and technology (S&T) capacity among NSP priority countries, which span all four levels defined by Wagner et al. (2001) — from scientifically advanced (e.g., Australia and Singapore) to lagging (e.g., Vietnam and Indonesia). In light of this, we aim to investigate country-level variations to gain a more granular understanding of the policy effects and the funding mechanisms driving the NSP initiative in science and technology cooperation.

References

Adams, J. (2013). The fourth age of research. *Nature*, *497*(7451), 557–560. https://doi.org/10.1038/497557a

Glänzel, W., Schubert, A., & -J. Czerwon, H. (1999). A bibliometric analysis of international scientific cooperation of the European Union (1985–1995). *Scientometrics*, *45*(2), 185–202. https://doi.org/10.1007/bf02458432

Gök, A., & Karaulova, M. (2023). How "international" is international research collaboration? *Journal of the Association for Information Science and Technology*, *75*(2), 97–114. https://doi.org/10.1002/asi.24842

Hottenrott, H., Rose, M. E., & Lawson, C. (2021). The rise of multiple institutional affiliations in academia. *Journal of the Association for Information Science and Technology*, *72*(8), 1039–1058. https://doi.org/10.1002/asi.24472

Karaulova, M., Gök, A., & Shapira, P. (2019). Identifying author heritage using surname data: An application for Russian surnames. *Journal of the Association for Information Science and Technology*, *70*(5), 488–498. https://doi.org/10.1002/asi.24104

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. https://doi.org/10.1016/s0048-7333(96)00917-1

Makkonen, T., & Mitze, T. (2016). Scientific collaboration between 'old' and 'new' member states: Did joining the European Union make a difference? *Scientometrics*, 106(3), 1193–1215. https://doi.org/10.1007/s11192-015-1824-y

Martin, B. R. (2016). R&D policy instruments – a critical review of what we do and don't know. *Industry and Innovation*, *23*(2), 157–176. https://doi.org/10.1080/13662716.2016.1146125

Miao, L., Larivière, V., Lee, B., Ahn, Y.-Y., & Sugimoto, C. R. (2024). Persistent hierarchy in contemporary international collaboration. https://doi.org/10.48550/ARXIV.2410.13020

Office of the President Republic of China (Taiwan) (2017). President Tsai's remarks at Yushan Forum: Asian Dialogue for Innovation and Progress. https://english.president.gov.tw/News/5232

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Simplemaps. (n.d.). *World Cities Database | Simplemaps.com*. Retrieved January 29, 2025, from https://simplemaps.com/data/world-cities

The Royal Society. (2010). *New frontiers in science diplomacy: Navigating the changing balance of power*. Science Policy Centre, The Royal Society. https://royalsociety.org/topics-policy/publications/2010/new-frontiers-science-diplomacy/

Wagner, C. S., Brahmakulam, I. T., Jackson, B. A., Wong, A., & Yoda, T. (2001). *Science & technology collaboration: Building capacity in developing countries?* RAND Corporation. https://www.rand.org/pubs/monograph_reports/MR1357z0.html

https://doi.org/10.51408/issi2025_113

Field Differences in External Funding: An Analysis of Funding Composition of Externally Funded Publications

Fredrik Niclas Piro[1], Henrik Karlstrøm[2], Ida Svege[3], Dag W. Aksnes[4]

[1]*fredrik.piro@nifu.no*, [2]*henrik.karlstrom@nifu.no*, [3]*ida.svege@nifu.no*, [4]*dag.w.aksnes@nifu.no*
NIFU – Nordic Institute for studies in Innovation, Research and Education, Økernveien 9, 0653 Oslo (Norway)

Abstract

The objective of this paper is to study field differences in external funding, using funding acknowledgment (FA) data from Web of Science for all papers with Norwegian authors in the years 2014-2022. Many studies use FA information as a dichotomous variable, but a large share of the FA information is institutional funding, and thus not external. To the best of our knowledge, we provide the first ever study of a large corpus of WoS publications where all FA information has been manually verified and standardized, and where institutional FA has been excluded. Our results indicate that using FA as a dichotomous variable overestimates the presence of external funding by 7.4 per cent. When only external funding is considered, we find that 58.9 per cent of all papers have external funding, and that this funding is highly unevenly spread across scientific fields. Furthermore, we find that external funding strongly increases the citation numbers of papers.

Introduction

In recent decades, there has been a steady growth in the magnitude of external funding of research globally (Tian et al., 2024; Heinze, 2008), i.e. shifting from institutional to project-based funding (Aagaard et al., 2021). Web of Science (WoS) allows for exploring this by using funding acknowledgements (FA) information. The purpose of this paper is to provide the first field comparison of external funding based on a large corpus of scientific publications, complementing previous funding studies in two ways. First, by clearly distinguishing between funding that is external and funding that is not, i.e. focusing only on external funding. This part of the analysis will provide a novel robustness check of the extent to which WoS funding acknowledgements data capture *external funding*. Second, by classifying all funding sources by type and country, thus exploring the roles of different funding types in different scientific fields. The analysis is based on all WoS publications in the period 2014-2022 with at least one Norwegian author: 259,198 papers with a total of 363,778 FAs listed.

Funding acknowledgments in Web of Science

The focus of this paper is on *publications*, and their funding (or lack thereof). The opportunity to analyze the presence of funding in WoS became possible in 2008 when FA data was introduced. Still, there are many caveats when using FA information. Once entered as FA in WoS, several functional challenges arise (Aagaard et al., 2021), of which most are attributed to the lack of standardization in how FA is reported. Most sources are listed by the funders' name (in the many ways it can be (mis-)spelled), whereas others are listed through project names (or

acronyms) or grant numbers. A major contribution from our study is the manual validation and standardization of all FA data listed in all publications. The FA text in WoS does not indicate what type of funding is provided[1], i.e. it does not differentiate between the two core types of funding, which is external funding and institutional funding (block funding). Hence, it is not possible to classify the large majority of reported FA by its many different funding types and purposes (El-Ouahi, 2024), i.e. whether it targets long-term or short-term projects (research or innovation), mobility, infrastructure, or being aimed towards particular groups of researchers, e.g. women or early-career researchers (Schweiger et al., 2024). The main difference between these two funding types is competition (by actively writing proposals or not), although this does come with some modifications as internal funding within an academic institution may be distributed following internal competition (Schweiger et al., 2024). Not all external funding, however, is based on competition (through submission of proposals to open calls). Foremost, industry funding (possibly also charity funding) may be channeled through researchers without competition and without a peer-review style assessment of proposals (Thelwall et al., 2023). We may also reason that much public funding from e.g. ministries may be provided from other types of processes than open calls.

Our operationalization of 'external funding' contrasts most of the literature on research funding which has either limited its focus to single programs (such as an excellence scheme or a specific call) or to more encompassing analyses based on 'everything' which is listed as FA. We believe such an approach is intertwined with the difficulty in distinguishing between 'funded' and 'unfunded' research (Thelwall et al., 2023), because a substantial part of WoS papers that do not recognize funding, are in fact funded by someone (in most cases, the researchers' institutions), whereas a substantial part of the 'funded' research, that is papers with FA are in fact listing institutional funding. We see 'external funding' as funding that is a) not institutional block funding, b) that is limited in time, and c) (mostly) obtained in open competition. This means that we are targeting funding that is channeled within a principal-agency framework (Gläser & Velarde, 2018), and with the funder of the research in a position to exercise influence on the content of the research carried out (Thelwall et al., 2023).

Several recent studies have used FA information either to compare funding across fields, or to classify FA data to show the engagement of different funding sources. For example, *Morillo* (2014) studied funding types (national/international) in papers from Spanish author addresses in four disciplines, revealing large differences across fields in the presence of FA and that international funding was associated with higher citation rates. In another Spanish study, *Alvarez-Bornstein, Diaz-Faes & Bordons* (2019) compared the funding patterns (public/private and national/international) of two medical fields. Here 89.9% of papers in virology had FA compared to 45% in Cardiac and Cardiovascular Systems. *El-Ouahi* (2024) studied publications with authors from the Middle East and North Africa (MENA), finding that about half of

[1] Infrequently, the FA text disclose types of funding, such as "project grant", "postdoc grant", "Professor Chair", "Endowment", "center of excellence funding", etc.

the papers had FA identical to the same organization as one of the authors, indicating institutional funding rather than external funding. *Diaz-Faes & Bordons* (2014) used FA as a dichotomous variable, but the results of this Spanish study are still relevant to us, because of its encompassing presentation of results across all fields in WoS. Here, FA was reported in two thirds of all papers, but with strong variations across fields. Tian et al. (2024) studied 13 million papers in WoS (2011-2020), with the aim of exploring changes in *universality* and *multiplicity* of funding over time. The former points to the presence of funding or not (a dichotomous approach), and the latter to the number of funders acknowledged. From 2011 to 2020 there was an increase in universality from 66.3% to 74.3%, and in multiplicity from 2.82 to 3.26 funders.

Data & methods

For this study, we applied a local version of WoS maintained by the Norwegian Agency for Shared Services in Education and Research. We retrieved 'Funding Agency' and 'Grant Number' fields in WoS, for all papers published in the years 2014-2022 with at least one Norwegian author. This dataset covers a total of 259,198 papers classified as original research papers, reviews and proceeding papers. All FA information has been manually read, interpreted (for example by internet searches) to identify the name, country and type of funding organization, as we are interested not only in whether there is a presence of external funding, but also in the *composition* of the funding. This requires a classification of *all* listed (external) funders. We have classified the funding organizations in the following categories: *Public sector* (which includes large programs such as the European Framework Programs for Research and Innovation), *Private sector* (which has been divided into three groups: pharmaceutical companies; companies operating within the oil and gas industry; other private companies), *Charity* (which includes non-governmental and non-profit foundations), *Other* (which includes organizations that do not have funding as their primary target, such as medical associations), and *Unknown*. The latter represents FA data that we at the time of writing (January 2025) have not yet correctly classified and currently present in 6.3 per cent of the papers, and thus a possible source of error. Nevertheless, we stress that our sample of FA is based on a correct classification of 93.7 per cent of all reported FA.

The sample of funders contain 1,756 unique Norwegian and 7,557 unique non-Norwegian funding sources. Considering the number of listed funding sources in the whole dataset we find 97,153 Norwegian funding acknowledgements and 266,625 non-Norwegian funding acknowledgements. Hence, in a study of Norwegian papers, only 26.7 per cent of the funding acknowledgements are Norwegian. The Research Council of Norway stands out as the most frequent single funding organization to papers with Norwegian authors (57,274 papers), followed by the European Framework Programs for Research and Innovation (EU FPs)[2] (20,125 papers), the Natural Science Foundation of China (NSFC) (5,351 papers), and the US agencies

[2] Please note that the second largest funding agency, the EU FPs is not equal to funding from the EU. It only points at the Framework Programs (FP6, FP7 and Horizon Europe), whereas other EU funding has been assigned other categories.

National Institutes of Health (NIH) (5,189 papers) and National Science Foundation (NSF) (4,437 papers). In comparing FA and external funding across scientific fields, we have grouped the papers based on their WoS journal categories into sixteen broad subject fields, following the classification suggested by NordForsk (2017). For papers in multidisciplinary journals and papers with missing information about journal categories, we have used the WoS macro, meso and micro topic classification scheme to discretionary regroup the papers to NordForsk's categories. Our study also includes a brief citation analysis. Here we calculated mean normalized citation scores (MNCS), where citation numbers are normalized by subject field, article type and year as well as a citation percentile indicator, identifying the top 10 percentile publications.

Results

In table 1 we show percentages of papers across subject fields that have reported FA, followed by percentages after exclusion of FA information that we consider to be institutional funding. At the overall level, we find that there is an overestimation of external funding equal to 7.4 per cent in the FA information in WoS. 63.9 per cent of the papers reported FA, but according to our classification, the percentage of papers with external funding is lower: 58.9 per cent. There are strong differences across subject fields in the presence of external funding. Fields from natural sciences are mostly above 70 per cent (exceptions being Engineering and Mathematics & Statistics at 57-58 per cent). Medical related fields show a gradient from Psychology (41.9%), Health sciences (51.9%), Clinical Medicine (59.6%) to Biomedicine & Molecular Biosciences (73.8%). Humanities (22.1%) and Social Sciences (37.2%) have the lowest shares.

Public funding sources account for the majority of FA and was reported in 53.3 per cent of all papers (Table 2), with the highest rates in Biology, Chemistry, Physics and Geosciences. In terms of being a complementary source of funding, charities are highly present in some of the subject fields where the public funding is lower than the overall percentage for public funding. This is foremost visible in Clinical medicine; Health sciences; and Psychology, where there are lower shares of papers with public funding, than for public funding overall.

Table 1. Percentage of papers reporting FA, and percentage of papers with external funding.

Subject field	Papers (n)	% with FA	% with external funding	% Overestimation
Agriculture. Fisheries & Forestry	9181	77.4	73.3	5.55
Biology	13285	80.6	77.1	4.44
Biomedicine & Molecular Biosciences	24562	78.1	73.8	5.78
Business Studies & Economics	8707	39.6	36.4	8.84
Chemistry	8257	80.0	75.4	6.07
Clinical medicine	35610	65.4	59.6	9.87

Computer & Information Science	12405	49.9	46.3	7.82
Engineering	34621	60.5	57.2	5.82
Geosciences	24104	77.9	74.8	4.21
Health sciences	19440	60.8	51.9	17.11
Humanities	8564	25.6	22.1	15.77
Materials science	6738	75.5	71.6	5.44
Mathematics & Statistics	4957	62.0	58.1	6.69
Physics	14209	77.0	74.4	3.43
Psychology	6823	49.2	41.9	17.37
Social sciences	27574	40.9	37.2	9.94
Total	259198	63.3	58.9	7.43

By contrast, here we find some of the most active involvement from charities: 28.6 per cent of papers in Clinical medicine; 16 per cent of papers in Health sciences and 10.5 per cent of papers in Psychology reported funding from charities. Nevertheless, the highest degree of funding from charities is reported in Biomedicine & Molecular Biosciences (28.7 per cent of the papers). Pharmaceutical companies were involved in 6.3 per cent of papers in Clinical medicine and 3.2 per cent of papers in Biomedicine & Molecular Biosciences; and oil/gas companies were involved in 5.2 per cent of papers in Geosciences and 4.1 per cent of papers in Engineering. Nevertheless, all three types of private funding display low percentages, i.e., they did not fund a large share of Norwegian papers.

Table 2. Percentage of papers with funding from key sources.

	Public	Charity	Private	Pharma	Oil	Other
Agriculture. Fisheries & Forestry	67.6	9.5	7.2	0.4	0.4	3.0
Biology	72.5	17.6	3.1	0.3	1.9	5.4
Biomedicine & Molecular Biosciences	67.0	28.7	3.1	3.2	0.6	3.3
Business Studies & Economics	33.0	3.7	0.7	0.1	0.3	0.6
Chemistry	72.5	9.3	4.3	0.3	2.7	1.8
Clinical medicine	45.7	28.6	3.4	6.3	0.1	3.4
Computer & Information Science	44.4	3.3	2.0	0.0	1.0	0.4
Engineering	53.9	2.7	4.8	0.1	4.1	1.1
Geosciences	70.4	8.8	2.6	0.1	5.2	3.5
Health sciences	42.3	16.0	2.5	1.0	0.1	3.2
Humanities	19.7	3.3	0.3	0.0	0.0	0.7
Materials science	68.7	5.8	6.2	0.0	1.8	0.9
Mathematics & Statistics	55.5	8.5	0.8	0.1	1.7	1.2
Physics	72.5	14.9	3.0	0.1	1.6	4.3
Psychology	36.5	10.5	0.7	0.3	0.1	2.1
Social sciences	34.2	3.9	0.7	0.0	0.3	1.1
Total	53.3	12.8	3.0	1.3	1.6	2.5

The different types of funders display different citation numbers across fields (not shown in tables). For example, for highly cited papers (within the top 10 per cent most cited from the same year and field), public funding varies from 8.1 per cent highly cited papers in Chemistry to 20.4 per cent in Clinical medicine. All funding types except the oil sector (9.6 per cent) have higher shares of highly cited papers than the world average (highest for Pharma and Charity; 28.8 and 19.9 per cent respectively). In table 3 we show differences in MNCS and highly cited papers for externally funded papers and papers without funding (which includes institutional FA). Presence of external funding is strongly associated with higher citation rates compared to papers without such funding (Table 3). On average externally funded papers have 45.3 per cent higher shares of highly cited papers and 35.5 per cent higher mean citation scores.

Table 3. Percentage of papers with funding from key sources.

	Per cent highly cited papers (10pctile)			MNCS (mean)		
	No funding	External funding	% diff.	No funding	External funding	% diff.
Agriculture. Fisheries & Forestry	10.2	13.5	32.7	1.10	1.38	25.6
Biology	11.8	14.8	25.8	1.11	1.40	25.5
Biomedicine & Molecular Biosciences	11.0	16.0	45.3	1.18	1.56	33.1
Business Studies & Economics	10.8	16.7	55.1	1.14	1.53	34.3
Chemistry	6.1	8.0	32.8	0.82	0.99	21.3
Clinical medicine	13.2	20.4	54.3	1.50	2.20	46.3
Computer & Information Science	10.3	13.7	33.0	1.04	1.36	30.3
Engineering	10.3	12.3	19.4	1.04	1.25	20.4
Geosciences	11.7	15.2	30.3	1.12	1.48	31.9
Health sciences	10.6	13.4	27.0	1.18	1.35	14.4
Humanities	9.7	22.1	28.0	1.12	2.27	102.3
Materials science	7.4	8.8	18.0	0.90	0.99	10.4
Mathematics & Statistics	7.6	11.5	51.4	0.89	1.26	42.0
Physics	9.1	13.3	46.1	0.97	1.45	49.5
Psychology	10.7	13.8	29.1	1.14	1.34	17.2
Social sciences	10.3	18.5	80.4	1.16	1.74	49.4
Total	10.7	15.6	45.3	1.18	1.59	35.5

The largest differences in citation indexes between funded and unfunded papers are seen in Humanities, which is a bit of special case due to this field's publishing and citation patterns. In other fields, compared to the total numbers, there is an especially strong effect of external funding in Social sciences; Clinical Medicine; Mathematics & Statistics; and in Physics (i.e. the difference in percentage between funded and unfunded papers are higher than the average for *both* highly cited papers

and MNCS). Note that in all fields, citation scores are higher for externally funded papers.

Discussion & conclusions

Our study has quantified the degree to which WoS FA data captures external funding and shown how such funding influence citation numbers across fields. We acknowledge that Norway is not representative to the world, given the high concentration of national funding through the Research Council of Norway, and with a much smaller representation of private foundations than for example in neighboring countries Sweden and Denmark. Still, the presence of international co-authorship, thus also international funding, is high in the papers we have studied. Being research in process, more work still needs to be done on the classification of (yet) *Unknown* FA sources. Later analysis will incorporate the aspect of *intensity* of funding and the interplay of different funding organizations (Tian et al., 2024), as for example EU publications display extremely high citation scores (Morillo, 2014). Nevertheless, the current analysis represents a novel contribution to the understanding of what WoS' FA data tells us, and how external research funding varies by field and how it is cited.

Acknowledgments

This work was supported by the Research Council Norway (RCN) [grant number 256223] (R-QUEST).

References

Aagaard, K., Mongeon, P., Ramos-Vielba, I. & Thomas, D.A. (2021). Getting to the bottom of research funding: Acknowledging the complexity of funding dynamics. *PLoS ONE,* 16(5), e0251488.

Alvarez-Bornstein, B., Diaz-Faes, A.A. & Bordons, M. (2019). What characterises funded biomedical research? Evidence from a basic and clinical domain. *Scientometrics,* 119(2), 805-825.

El-Ouahi, J. (2024). Research funding in the Middle East and North Africa: analyses of acknowledgements in scientific publications indexed in the Web of Science (2008-2021). *Scientometrics*, 129, 2933-2968.

Dias-Faez, A.A. & Bordons, M. (2014). Acknowledgements in scientific publications: Presence in Spanish science and text patterns across disciplines. *Journal of the Association for Information Science and Technology,* 65(9), 1834-1849.

Heinze, T. (2008). How to sponsor ground-breaking research: a comparison of funding schemes. *Science and Public Policy*, 35(5), 302-318.

Liu, W., Tang, L. & Hu, G. (2020). Funding information in Web of Science: an updated overview. *Scientometrics,* 122, 1509-1524.

Morillo, F. (2019). Collaboration and impact of research in different disciplines with international funding (from the EU and other foreign sources). *Scientometrics*, 120, 807-823.

NordForsk (2017). *Comparing Research at Nordic Higher Education Institutions Using Bibliometric Indicators. Covering the Years 1999-2014.* Policy Brief 4/2017. Oslo: NordForsk.

Schweiger, G., Barnett, A., van den Besselaar, P., Bornmann, L., De Block, A., Ioannidis, J.P.A., Sandström, U. & Conix, S. (2024). *The Costs of Competition in Distributing Scarce Research Funds*. arXiv:2403.169341v1 (25.03.2024).

Tian, W., Cai, R., Fang, Z., Xie, Q., Hu, Z. & Wang, X. (2024). Research funding in different SCI disciplines: A comparison based on Web of Science. *Quantitative Science Studies*, 5(3), 757-777.

Thelwall, M., Simrick, S., Viney, I. & van den Besselaar, P. (2023). What is research funding, how does it influence research, and how is it recorded? Key dimensions of variation. *Scientometrics*, 128, 6085-6106.

https://doi.org/10.51408/issi2025_146

Geographies of Underrecognition: Citation Disparities in Russian Studies

Katerina Guba[1], Elena Chechik[2], Angelika O. Tsivinskaya[3], Artur Pecherskikh[4], Nikita Buravoy[5]

[1]*kguba@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[2]*echechik@eu.spb.ru*
Europa-Universität Flensburg, Auf dem Campus 1, Flensburg (Germany)

[3]*atsivinskaya@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[4]*apecherskikh@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[5]*solvelimits@gmail.com*
Independent Researcher

Abstract

This paper focuses on the global inequalities in academic recognition within the field of Russian Studies, focusing on the geographical dimension of citation bias. Historically, Russian Studies has been shaped primarily by Western institutions, with limited contributions from local scholars during the communist era. Despite increased participation by Russian academics in international scholarship, citation disparities persist, reflecting broader systemic inequalities in global knowledge production. Using a large dataset of publications and citations, we analyze whether an author's country of affiliation influences citation rates, specifically examining whether papers by Russian-affiliated scholars are cited less frequently than those from other regions. Our findings align with previous research demonstrating that peripheral regions, including Russia, are consistently undercited compared to core academic hubs like North America and Europe.

Introduction

Citation bias has been extensively studied, primarily with a focus on gender and racial disparities (Dion et al., 2018). At the country level, citation bias manifests as a tendency among researchers to preferentially cite studies authored by Western scholars. This bias reinforces the overrepresentation of mainstream findings in the scientific literature as studies from non-Western or peripheral contexts may be neglected or underrepresented, perpetuating systemic inequalities in global knowledge production. Our study seeks to examine the factors that explain citation disparities, with a particular focus on the geographical dimension of citation bias. By emphasizing global inequalities in knowledge production, we aim to contribute to the understanding of how geographic factors shape the visibility and recognition of scholarly work (Qiu et al., 2025; Gomez et al., 2022).

Studies on citation inequality assume that all analyzed publications are from prestigious journals, meaning that there are no systematic variations in paper quality based on the author's country. This implies that factors beyond the quality of the paper, such as institutional or regional biases, influence citation patterns (Sin, 2011). However, arguments about citation inequality must also account for topic specialization, which can vary by country. Countries may specialize in topics with fewer active researchers, thereby influencing the number of citations their papers receive (Gomez et al., 2022). Using Russian Studies as a case, this paper seeks to account for topic specialization across different national contexts. Unequal recognition may stem from differences in the geographic focus of researchers based in core and peripheral countries. Scholars often gravitate toward familiar objects of study, and Western countries are privileged as the primary focus of academic research (Krause, 2021). As a result, studies focusing on peripheral regions are likely undercited, partly because fewer researchers engage with these topics. Moreover, when such studies are available, scholars often prefer to cite work with a geographic, economic, or social focus that aligns with their own research, bypassing studies on less familiar regions. In this study, we address this issue by investigating whether the country of an author's institutional affiliation affects the citation rate of manuscripts. We limit our scope to papers focused on a single geographic region—Russia—to eliminate variance in geographic scope as a factor influencing citation patterns.

This study relies on scientometric tools and a comprehensive bibliographic database—Web of Science (WoS)—to collect journal articles focusing on Russia in the social sciences over a 30-year period (1990–2020). While electronic databases provide access to extensive information for studying knowledge production, certain database-specific limitations can pose challenges for interdisciplinary fields like Russian Studies and post-Soviet area studies. To overcome these challenges, we developed a sophisticated search query designed to capture a broad range of relevant literature.

Material and methods

This study employs bibliometric analysis of publications in Russian Studies indexed in the Web of Science (WoS) database over a 30-year period (1990-2020). The dataset comprises 29,826 journal articles in the social sciences, identified using an advanced keyword search strategy. To create the main dataset, we employed diverse bibliometric methods for the identification of papers with a focus on Russia. The process of data collection included seeding a pilot dataset for keywords, selection of keywords, storing the primary dataset, selection of papers by experts and the cleaning of affiliation information. The use of around 1,271 keywords relevant to Russia resulted in 29,826 articles stored on the Web of Science database for the period 1990 – 2020 (the list is available in (Guba et al., 2024). We use the list of keywords to get all academic papers written in English during the period 1990 – 2020. To be stored, a paper has to contain at least one word from keywords in titles, abstracts or keywords. Our initial WoS query yielded in 55 709 (the database was queried in January, 2022), only article and review were taken into account. Since this list is likely to contain redundant papers, additional steps were needed to provide a corpus of articles appropriate for further analysis.

Our next stage was to resort to expert assessments once again to narrow down the dataset leaving only relevant publications. This step was necessary as querying articles by keywords might result in partially or completely unrelated documents. Since articles containing Russia in their title can be treated suitable with a substantial degree of certainty, such papers were not subject to expert assessment and immediately marked relevant. Thus, four experts received a shortened dataset of 40,647 papers to be checked for compliance with the topic. They read titles and examined keywords and abstracts. For the whole coded dataset, agreement and partial agreement constituted approximately 68.5% and 96.8%, respectively. Overall, an article was accepted if it contained the substring Russia in its title or if at least three out of four experts marked it as 1 (related). 29,826 papers (roughly 54% of the whole corpus) met this criterion.

For the citation analysis in this study, we selected only 15,078 publications indexed in the Social Sciences Citation Index, as citation analysis has significant challenges in the humanities.

Results

Given that the outcome variable, citations received, is not normally distributed, instead of using raw citation counts, we rely on the Mean Normalized Citation Score (MNCS), which represents the average number of citations for publications normalized by research field and publication year. This indicator reflects how a publication's citation performance compares with the global average. For the regression analysis, we binarize the variable, with 1 representing citations above the world average (MNCS > 1) and 0 representing citations below or equal to the world average (MNCS ≤ 1).

Our aim is to test whether the citations received are related to the author's geographic affiliation (in terms of the country), which is the main focus of this study. We coded authors' geographic affiliations using the information about their country of employment, as indicated by the correspondence address, rather than their nationality. This approach is widely used in scientometric research to draw conclusions about the country with which an author is affiliated. For authors with multiple affiliations, only the first affiliation was considered. For multi-authored papers, the total author counting method was employed, whereby data for all contributing authors were coded. Finally, we categorized the countries into several subregions based on the classification provided by the United Nations Statistics Division, with some adaptations to account for our focus on Russian scholars and the low number of articles in certain regions. The subregions include North America, Russia, Eastern Europe, Northern Europe, Southern Europe, Western Europe, Oceania, and Asia. North America accounting for the largest proportion of authors (40.75%), followed by Northern Europe (21.96%) and Russia (16.02%). Smaller contributions are observed from Asia (5.75%), Western Europe (8.71%), Eastern Europe (2.76%), Oceania (2.15%), and Southern Europe (1.90%), for a total of 17,284 articles.

The key step in studying the relationship between article citability and geographical factors is to control for the prestige level of the publishing journal (Abramo et al., 2024). In this study, we used SJR, or the SCImago Journal Rank indicator, as the

metric for journal impact. SJR ranks scholarly journals based on citation weighting schemes and eigenvector centrality accounting for the visibility of journals citing a given journal's set of papers. Based on findings from previous research, more variables were included to account for variance in citations received that are related to authorship patterns. Regarding, the coauthorship type, international collaborations tend to be cited more frequently, a trend confirmed by both cross-national analyses and case studies of specific countries (see Olechnicka et al. (2019) for a review). In addition, the number of authors is related to a larger number of citations (Sin, 2011). The year of publication was included, as articles that are published earlier tend to have more time to accumulate citations, but there may also be aging of older articles (Sin 2011). Regarding the document type, we limited our analysis only to journal articles.

In summary, this study tested seven variables: (1) author's subregion, (2) journal SJR, (3) authorship type (4) number of authors, and (5) publication year. This research does not aim to build a full model for citation count prediction given the complexity of phenomenon as researchers found a range possible factors (Abramo et al., 2024). Rather the current logistic regression analysis aims to test whether geographical factors are, indeed, related to significant different citation counts.

The value of MNCS for Northern American publications is 1.2; the MNCS score for the European articles is 1.1 in 1990-2020 with observed differences between different parts of European regions. The MNCS for Russia in 1990-2010 was 0.6, while for the period 2010-2020 the value was 0.9 meaning that Russian articles started to receive almost the same number of citations as on average in the world in the same research field and publication year.

A logistic regression analysis was conducted to test the relationships between the geographical factor and other variables with the likelihood of an article being cited above the world average (Table 1). The odds ratio (OR) was used to evaluate how each variable affected the outcome variable with an OR greater than 1 indicating that articles with a given characteristic are more likely to be cited above the world average.

Table 1. Logistic regression analysis.

Variables	*Odds Ratio*
Dependent variable: Citation above world average	
Journal citation metric SJR	1.988***
	(0.0538)
Collaboration (reference category – solo collaboration)	
International collaboration	1.821***
	(0.116)
National collaboration	1.441***
	(0.0805)
Number of authors	1.033
	(0.0206)

Region (reference category –North America)	
Asia	0.790***
	(0.0623)
Russia	0.630***
	(0.0363)
East Europe	0.742**
	(0.0886)
North Europe	1.067
	(0.0546)
South Europe	0.684***
	(0.0913)
Western Europe	0.713***
	(0.0512)
Oceania	1.089
	(0.154)
Year	0.998
	(0.00320)
Constant cut1	0.0409
	(0.263)
Observations	13,058

*** p<0.01, ** p<0.05, * p<0.1

The logistic regression results revealed significant associations between the variables and the outcome of interest. The Scimago Journal Rank (SJR) demonstrated a strong positive effect with an odds ratio of 1.988 ($p < 0.01$), indicating that higher-ranked journals are significantly more likely to publish articles that achieve citation counts above the global average. As predicted, collaboration type also plays a crucial role: international collaboration yielded an odds ratio of 1.821 ($p < 0.01$), while national collaboration showed an odds ratio of 1.441 ($p < 0.01$), both indicating a positive effect compared to solo authorship. In contrast, the number of authors (OR = 1.033, $p > 0.05$) and publication year (OR = 0.998, $p > 0.05$) did not exhibit significant impacts on citation likelihood.

Regional effects, our primary focus, showed notable variability. Articles authored by researchers from Asia (OR = 0.790, $p < 0.01$) and Russia (OR = 0.630, $p < 0.01$) were less likely to achieve citation counts above the world average compared to the reference region (North America). Among European authors, papers from East Europe (OR = 0.742, $p < 0.01$) and South Europe (OR = 0.684, $p < 0.01$) also garnered fewer citations compared to those from North America. Interestingly, no significant differences were found between North American authors and those from North Europe or Oceania/Australia. For Russian articles, the probability of being cited above the world average is the lowest compared to other regions (e.g., 0.31), while for North America, Northern Europe, and Oceania, this probability is notably higher, ranging from 0.42 to 0.44. In other words, even when scholars affiliated with Russian institutions overcome the challenges of publishing in reputable journals, their papers tend to receive fewer citations. These findings align with previous

studies in other disciplines, which have shown that Russian scholars do not achieve the same influence from their published research as scholars from frontier regions (Dyachenko & Pisklyakov, 2010).

Discussions

Area studies, such as Russian Studies, occupy an interstitial epistemic space, serving both as a research subject and as a field where many scholars are geographically situated (Kaczmarska & Ortmann, 2021). Local researchers possess valuable knowledge of the empirical context and cultural experience, which often makes them more informed experts compared to foreign scholars. However, as they are positioned on the academic periphery, their chances of getting published and cited are unequal. In this paper, we focus on the issue of unequal recognition in knowledge production about Russia by analyzing the quantity and impact of academic publications.

We observed unequal citation recognition across countries and world regions. What might explain these disparities? One possibility is that journal metrics fail to capture systematic differences in citation potential, though similar results have been observed in studies of citation patterns for Chinese papers (Qiu et al., 2025). Another explanation relates to network effects (Dion et al., 2018): scientists may be less aware of research produced by Russian authors. To gain citations, authors require access to "the networks that provide broad exposure to research findings" (Qiu et al., 2025). Previous studies have identified a "home bias," where scientists disproportionately cite researchers from the same region, language, or nation (Pasterkamp et al., 2007; Sin, 2011; Qiu et al., 2025). Given the larger size of the Western scholarly community, it is predictable that their articles would have more chances of being cited. Conversely, publishing internationally remains a significant challenge for Russian scientists, meaning that there are fewer Russian scholars publishing in international journals, and consequently fewer opportunities for them to cite each other. Building robust academic networks is often contingent on significant international experience – an opportunity that many Russian scholars lack.

Citation counts alone fail to capture the sociological interpretaion underlying how scholars recognize the work of their peers, highlighting the need for a deeper analysis of citing behavior. At a minimum, we have gathered sufficient evidence to justify continuing this line of inquiry. The most promising results may be obtained through experimental surveys, which offer opportunities to test hypotheses about the social factors influencing citation behavior by presenting differently formulated questions to control and experimental groups. Studies using experimental designs have already demonstrated the existence of evaluation biases based on factors such as gender and institutional prestige (Knobloch-Westerwick et al., 2013).

Acknowledgments

The research was supported by Russian Science Foundation, Grant/Award Number 25-28-01490.

References

Abramo, G., D'Angelo, C. A., & Grilli, L. (2024). The role of non-scientific factors vis-à-vis the quality of publications in determining their scholarly impact. *Scientometrics*, *129*(8), 5003-5019.

Dion, M. L., et al. (2018). Gender and citation patterns in political science. *PS: Political Science & Politics*.

Gomez, C. J., Herman, A. C., & Parigi, P. (2022). Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour*, *6*(7), 919-929.

Guba, K., Chechik, E., Tsivinskaya, A. O., & Buravoy, N. (2024). Global Ranking of Expertise about Russia. *Problems of Post-Communism*, 1-11.

Kaczmarska, K., & Ortmann, S. (2021). IR theory and area studies: A plea for displaced knowledge about international politics. *Journal of International Relations and Development*, *24*(4), 820-847.

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science communication*, *35*(5), 603-625.

Krause, M. (2021). *Model cases: On canonical research objects and sites*. University of Chicago Press.

Qiu, S., Steinwender, C., & Azoulay, P. (2025). Who stands on the shoulders of Chinese (scientific) giants? Evidence from chemistry. *Research Policy*, *54*(1), 105147.

Olechnicka, A., Ploszaj, A., & Celińska-Janowicz, D. (2019). *The geography of scientific collaboration* (p. 236). Taylor & Francis.

Pislyakov, V., & Dyachenko, E. (2010). Citation expectations: are they realized? Study of the Matthew index for Russian papers published abroad. *Scientometrics*, *83*(3), 739-749.

Pasterkamp, G., Rotmans, J., de Kleijn, D., & Borst, C. (2007). Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics*, *70*(1), 153-165.

Sin, S. C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008. *Journal of the American Society for Information Science and Technology*, *62*(9), 1770-1783.

https://doi.org/10.51408/issi2025_100

Harnessing Data Papers: An Analysis of Their Role in Scientific Data Dissemination and Reuse

Liyue Chen[1], Xiaomin Liu[2]

[1] *chenliyue@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, 33 Beisihuan Xilu, Zhongguancun, Haidian District, Beijing (China)

[2] *liuxm@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, 33 Beisihuan Xilu, Zhongguancun, Haidian District, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, 80 East Zhongguancun Road, Haidian District, Beijing (China)

Abstract

This research in progress studies the scholarly role of data papers in scientific data sharing and reuse. Data papers are a crucial publishing form for scientific data, tasked with fully leveraging the value of data science. This study, based on extensive citation context data and large language models, investigates the actual contributions and specific citation purposes of data papers in citing documents. The results indicate that data papers indeed play a role in disseminating scientific data for reuse during scholarly communication, yet their potential has not been fully realized, with certain data papers still serving primarily as methodological support and overviews of data development backgrounds. The impact of data papers also varies across disciplines; fields such as life sciences, natural resources and environmental sciences place greater emphasis on the role of data papers, with richer integration and utilization of scientific data based on them. However, the value of data papers in humanities, social sciences, and some fundamental disciplines remains underexplored.

Introduction

The rapid transformation of scientific paradigms underscores the significance of scientific data. However, as scientific data rapidly accumulates and is widely shared, several issues have become increasingly evident: researchers' reluctance to share data (Gajbe et al., 2021; Mattern et al., 2024), the fragmentation of data resources (Shen et al., 2024), unclear data ownership (Sheng & Yuan, 2021), and challenges in controlling data quality. To address these issues and promote the reproducibility of research findings, the data paper has emerged as an important academic publication format. Data papers are peer-reviewed scholarly publications that provide a standardized description of scientific datasets (Carlson & Oda, 2018; Chavan & Penev, 2011). Typically, data papers detail the methods of data collection and processing, data structure and storage format, methods of data use, and access pathways (Kim, 2020), serving as a 'manual' for the data.

Existing research has analyzed the crucial role of data papers in promoting open data reuse from the perspective of the motivations for data sharing. The reluctance of researchers to share and reuse scientific data can generally be divided into two categories. On one hand, data producers lack the motivation to share, as researchers are uncertain whether their actions of sharing will be properly rewarded (Mattern et

al., 2024; Tenopir et al., 2015; Wallis et al., 2013). On the other hand, there is also a lack of sufficient production information to support data reuse (Borgman, 2012; Curty et al., 2017). Compared to other data publication formats, data papers secure the data producers' right to discovery priority and academic reputation through publication and formal citation. They also stimulate data sharing and reuse by controlling the quality of data through a rigorous peer-review process (Thorisson, 2009; Zhao et al., 2018).

A few studies have also explored the actual dissemination impact of data papers on the reuse of scientific data, starting from the citation practices of data papers. Jiao and Darch (2020) analyzed the citation context of 103 data papers in earth sciences and physics through manual interpretation and found that data papers have not fully realized their potential in promoting data reuse. Research on citation behaviors in the biomedical field shows a steady increase in formal citations of data papers for the purpose of data usage (Jiao et al., 2024). Similarly, in the humanities and social sciences, data papers have had a positive impact on the influence of datasets in related research papers (McGillivray et al., 2022). Overall, existing research on data papers is limited in scale, relies primarily on manual annotation for interpreting the scholarly communication role of data papers, and is only conducted within a few disciplines, which does not provide a comprehensive view of the development of data papers.

The purpose of this study is to explore the role of data papers in the open sharing and informed reuse of scientific data during the academic communication. Three research questions are considered: (1) Do data papers make a data-related contribution to the studies that cite them? (2) If they do data-related contribute, for what purposes do the citing works use the scientific data? (3) Does the role of data papers vary across different disciplines? To address these questions, this research conducts a citation context analysis on data papers across all disciplines, which includes identifying actual contributions and analyzing citation purposes, aiming to better understand the facilitative role of data papers in the sharing and reuse of scientific data.

Data and Methods

In our study, we limited the document type to 'data paper' within the Web of Science Core Collection, covering the period from 1980 to 2024. We retrieved a total of 17,318 data papers, of which 82.24% were cited, with an average citation count of 15.99 per paper. To categorize the disciplines of these data papers, we used the InCites citation topics (macro) schema to map the papers onto 10 disciplines. To analyze the citation context characteristics of these papers, we collected citation context data from citing documents via the *Scites* platform (https://scite.ai). Ultimately, 12,422 data papers were matched with 219,707 citation context entries.

The identification of the actual contributions of data papers employed the automatic recognition method proposed in our previous study (Chen et al., 2024), which classifies the contributions of papers into five categories: theoretical, experimental, methodological, data-based, and other. This classification is performed using a fine-tuned Llama2-13B large language model, which achieved an accuracy rate of 0.94.

To automatically identify the citation purposes of data papers, we utilized large language model techniques in our experiments. Building on existing research (Gregory et al., 2019), we categorized citation purposes into seven types: background, calculation, integration, verification, inspiration, and other. We tested various prompt schemes and different large models (including DeepSeek-R1 and GPT-4o), and ultimately determined that GPT-4o had the best recognition performance, with an F1-score of 0.81.

Results

In-text citation characteristics of data papers

We first analyzed the frequency of in-text mentions of data papers in citing documents. The frequency of in-text mentions for the data papers analyzed was 1.673, which is similar to that of traditional academic papers (Chen et al., 2022; Hsiao & Chen, 2018). In citing documents, instances where data papers were mentioned only once accounted for approximately 66.87%, while mentions two times or more accounted for 33.13%. When we conducted the analysis by different disciplines, we found that data papers in fields such as Physics, Earth Sciences, Agriculture, Environmental and Ecology had an average in-text mention frequency higher than that of all disciplines. In contrast, the Art and Humanities, Social Sciences, and Mathematics had lower in-text mention frequencies. This outcome reflects, to some extent, the varying degrees of emphasis placed on scientific data across different research areas.

We also analyzed the distribution of in-text locations for data papers within the citing studies. As shown in Figure 1, data papers are most frequently mentioned in the 'Materials and Methods' and 'Introduction' sections, accounting for 24.28% and 23.33% respectively. In contrast, they appear less frequently in the 'Results' and 'Discussion' sections. Compared to traditional academic papers, which are primarily mentioned in the 'Introduction' and 'Discussion' sections (Bertin et al., 2016; Voos & Dagaev, 1976), data papers indeed show a distinct characteristic of explicit data support.

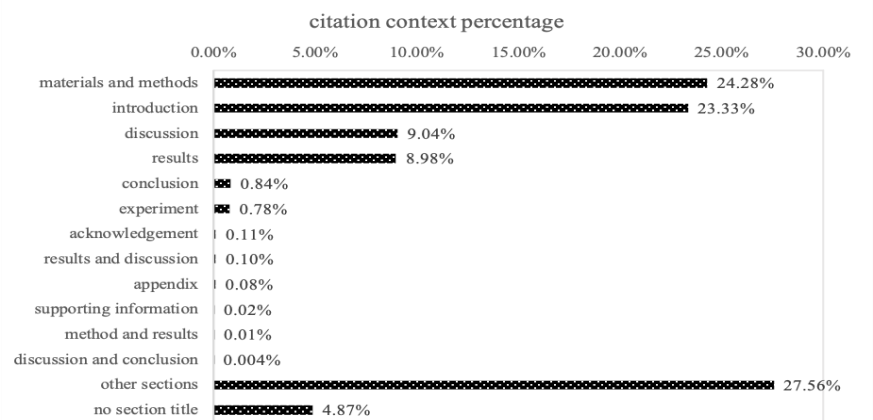


Figure 1. The distribution of in-text locations for data papers within the citing studies.

An analysis of the in-text locations of data papers across various disciplines reveals that Mathematics data papers are most frequently mentioned in the 'Materials and Methods' section, accounting for 42.54% of mentions (Figure 2). This is followed by Earth Sciences and Clinical & Life Sciences. In contrast, in fields such as Humanities and Social Sciences, Engineering and Materials Science, data papers are more often cited in the 'Introduction' section, seemingly serving more as a background overview. Moreover, compared to other fields, data papers in Chemistry and Clinical & Life Sciences are relatively more frequently cited in the 'Results' section. This trend may stem from the reliance of these disciplines on scientific experimental processes and experimental data, using cited scientific data from other studies for comparative analysis and validation of results in their current research.

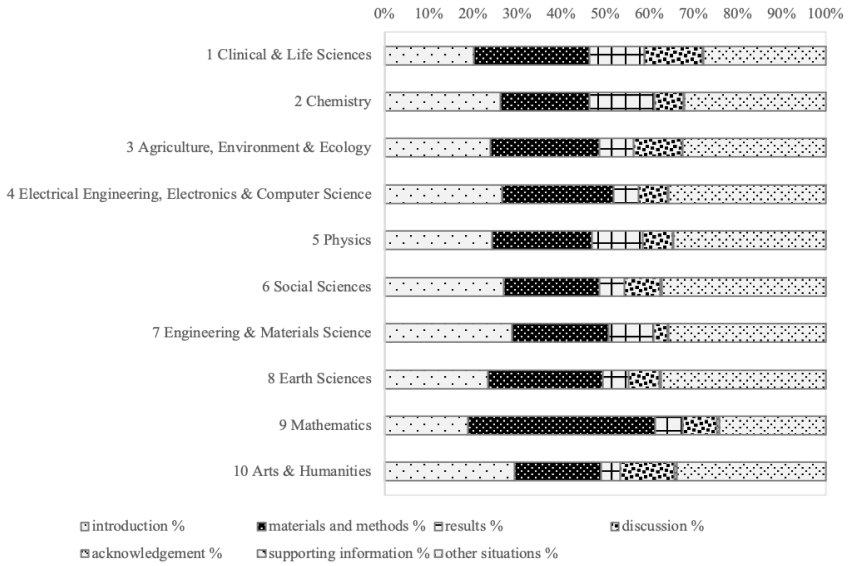


Figure 2. The distribution of in-text locations for data papers across 10 disciplines.

The actual contribution roles of data papers

By identifying the actual contribution types of data papers, it was found that data papers indeed make a data-based contribution to the citing research in 42.7% of cases. However, in more than half of the citation context instances, data papers play roles in other aspects, including providing methodological support for the citing research, being used for experimental comparison and result validation, and offering background information for the cited studies.

The actual contributions of data papers in various research areas also show significant differences (Figure 3). In fields such as Electrical Engineering, Earth Sciences, and Agronomy, the proportion of data-based contributions is relatively high, whereas in Engineering and Materials Science, Mathematics, and Chemistry, the proportion of data contributions is relatively lower. In Engineering and Materials Science, data papers are more focused on corroborating and supporting experimental results, while in Mathematics and Chemistry, data papers play a role in methodologies such as mathematical formulas and experimental schemes.

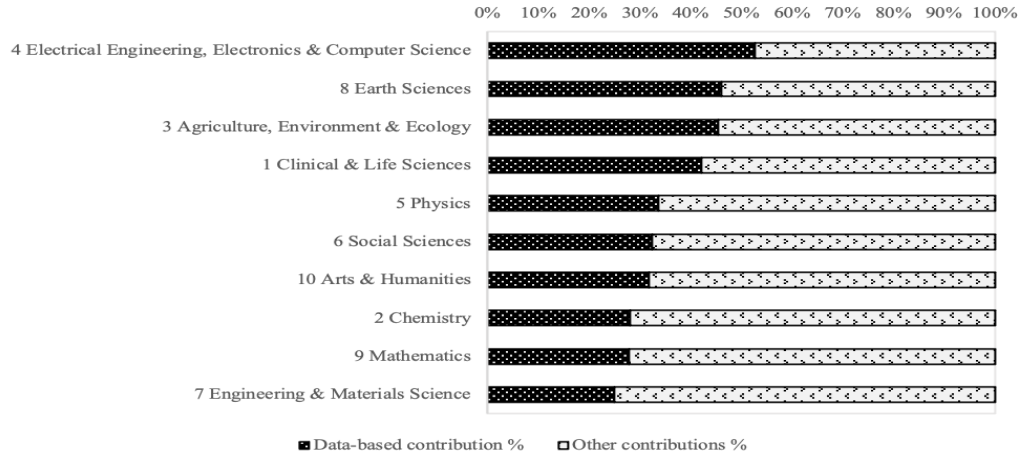


Figure 3. Distribution of actual contribution types made by data papers in 10 research areas.

The purpose of citing scientific data in data papers

The type of actual contribution reflects whether data papers have made a contribution related to citing works, focusing specifically on the data core of the cited papers. To further understand the purposes behind the citation of data papers in citing studies, we conducted a citation purpose analysis on the contexts where data papers clearly have a data-based contribution. As Figure 4 shows, in nearly 50% of cases, although data papers provide a data-based contribution, they are merely mentioned by the citing studies to elaborate on the background knowledge of the research. Secondly, a certain proportion of citations in citing studies are due to their use in experimental calculations, accounting for 15.19%. Additionally, citing papers use the cited literature for data integration, comparative data analysis, and data benchmarking and validation, with these three citation purposes having similar distributions. Very few data papers serve as a source of research inspiration for the citing literature.

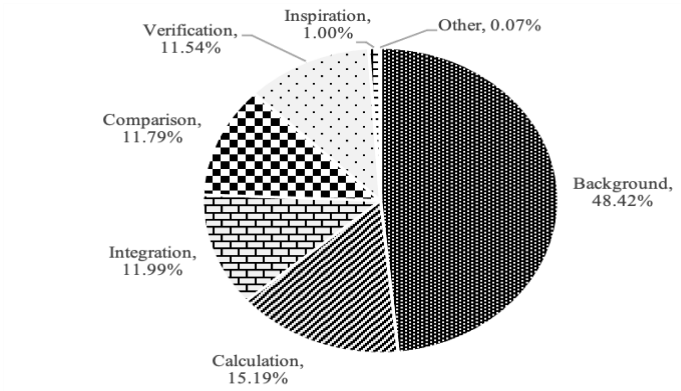


Figure 4. Citation purposes distribution of data papers with data contributions.

Figure 5 shows that the distribution of citation purposes in cited data papers is relatively similar across various disciplines. Data papers in Life Sciences and Earth Sciences have a higher proportion of citations for usage purposes (including calculation, comparison, integration and verification) compared to the average across all disciplines. In contrast, these usage purposes are relatively lower in Art and Humanities, as well as Mathematics.

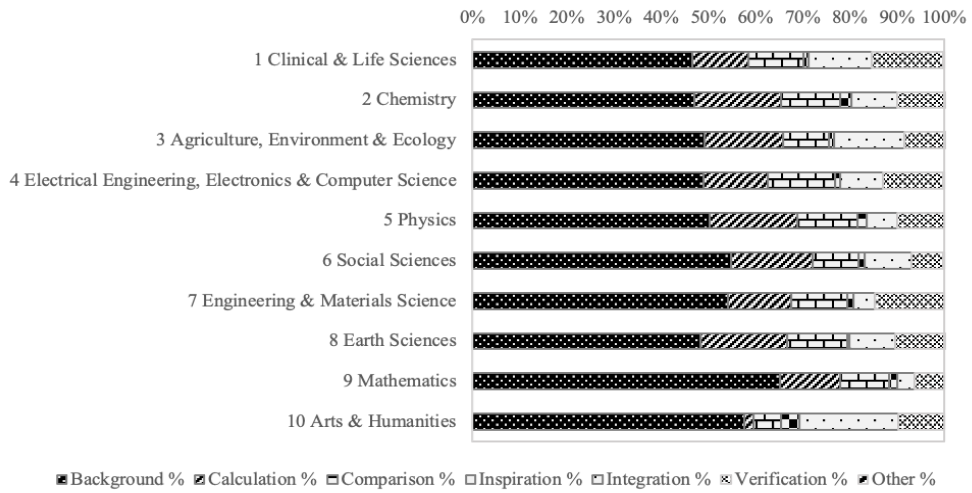


Figure 5. Citation purposes distribution of data papers with data contributions (10 disciplines).

Discussion and Conclusion

Data papers not only promote transparency and reproducibility in scientific research but also confer academic credit to data producers. It is very meaningful to explore how data papers can help scientific data fully realize its innovative value.

Our study finds that data papers indeed play a clear role in the dissemination and reuse of scientific data, yet there is substantial space for improvement. A significant number of citations to data papers still stem from methodological support, experimental comparisons, and result validation, or describing the current state of data development relevant to the research questions. By comparing citation context characteristics of data papers across various research areas, we observe differences in attention and usage levels towards data papers among disciplines. Areas such as Clinical & Life Sciences, Agriculture, Environment & Ecology are experiencing rapid development in scientific data and data papers, with a more pronounced trend in data usage based on citations to data papers. However, fields like the Arts & Humanities, and Mathematics have smaller volumes of data papers, with citations primarily focusing on confirming research viewpoints and describing relevant backgrounds. Our subsequent research questions include: (1) By conducting annual statistics, we will explore whether the role or function of data papers has changed. (2) From the perspective of academic publishing standards, we will further investigate the publishing attributes of cited data papers and the citing literature.

References

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology, 67*(1), 164-177.

Borgman, C. L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078.

Carlson, D., & Oda, T. (2018). Data publication - goals, practices and recommendations. *Earth System Science Data, 10*(4), 2275-2278.

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *Bmc Bioinformatics, 12*, S2.

Chen, L., Ding, J., Song, D., & Qu, Z. (2024). Exploring Scientific Contributions through Citation Context and Division of Labor. *arXiv* preprint arXiv:2410.13133.

Chen, L. Y., Ding, J. L., & Lariviere, V. (2022). Measuring the citation context of national self-references. *Journal of the Association for Information Science and Technology, 73*(5), 671-686.

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *Plos One, 12*(12), e0189288.

Gajbe, S. B., Tiwari, A., Gopalji, & Singh, R. K. (2021). Evaluation and analysis of Data Management Plan tools: A parametric approach. *Information processing & management, 58*(3),102480.

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A. and Wyatt, S. (2019). Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *Journal of the Association for Information Science and Technology, 70*: 419-432.

Hsiao, T. M., & Chen, K. H. (2018). How authors cite references? A study of characteristics of in-text citations. *Proceedings of the Association for Information Science and Technology, 55*(1), 179-187.

Jiao, C., & Darch, P. T. (2020). The role of the data paper in scholarly communication. *Proceedings of the Association for Information Science and Technology, 57*(1), e316.

Jiao, H., Qiu, Y. H., Ma, X. W., & Yang, B. (2024). Dissemination effect of data papers on scientific datasets. *Journal of the Association for Information Science and Technology, 75*(2), 115-131.

Kim, J. (2020). An analysis of data paper templates and guidelines: types of contextual information described by data journals. *Science Editing, 7*(1), 16-23.

Mattern, J. B., Kohlburn, J., & Moulaison-Sandy, H. (2024). Why academics under-share research data: A social relational theory. *Journal of the Association for Information Science and Technology, 75*(9), 988-1001.

McGillivray, B., Marongiu, P., Pedrazzini, N., Ribary, M., Wigdorowitz, M., & Zordan, E. (2022). Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences. *Publications, 10*(4), 39.

Shen, Z.H., Zhu, X.J., Wang, H.J., et al. (2024). Research data network: concept, systems and applications. *Frontiers of Data & Computing, 6*(4), 3-21.

Sheng, X.P., & Yuan, Y. (2021). Data rights governance in open sharing of scientific data. *Journal of Library Science in China,* (5), 80-96.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *Plos One, 10*(8), e0134826.

Thorisson, G. A. (2009). Accreditation and attribution in data sharing. *Nature Biotechnology, 27*(11), 984-985.

Voos, H., & Dagaev, K. S. (1976). Are All Citations Equal? Or, Did We Op. Cit. Your Idem? *Journal of Academic Librarianship, 1*(6), 19-21.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *Plos One, 8*(7), e67332.

Zhao, M., Yan, E., & Li, K. (2018). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology, 69*(1), 32-46.

https://doi.org/10.51408/issi2025_111

Exploring the Effects of Migration for Social and Humanity Researchers (Research in Progress)

Alena Nefedova[1], Elizaveta Chefanova[2]

[1]*anefedova@hse.ru*, [2]*echefanova@hse.ru*
Institute for Statistical Studies and Economics of Knowledge,
National Research University Higher School of Economics, Myanitskaya st., 11,
Moscow (Russian Federation)

Abstract

The international mobility of researchers is a central topic in global academic discourse due to its complex and multifaceted consequences. While mobility can provide opportunities for knowledge acquisition, professional growth, and access to advanced resources, it also introduces challenges such as career instability, loss of social networks, and barriers to integration into host academic environments. This study examines these dynamics with a focus on researchers in the humanities and social sciences, employing the theory of three researcher careers (Gläser & Laudel, 2015) as a conceptual framework. The model highlights three dimensions of career development—cognitive, organizational, and community—which are uniquely affected by mobility. Based on 20 in-depth semi-structured interviews with humanities and social science researchers who experienced long-term international mobility, the findings underscore the dual nature of mobility. While it offers professional growth and access to global resources, it also disrupts career trajectories, particularly in disciplines with less international standardization or commercial applicability.

Preliminary results show that language plays a pivotal role in determining migration destinations. Researchers frequently select countries where they can work in their native or familiar languages. Others adapt by taking roles outside their research fields, such as teaching Russian as a foreign language or similar positions in high demand internationally. However, such roles often limit their ability to advance their cognitive careers.

The study also highlights the emotional and professional toll of constant migration. Researchers report exhaustion from navigating unstable employment conditions and the challenges of rebuilding professional networks in host countries. This disruption diminishes their community career standing, as they lose the professional connections that previously facilitated access to resources and opportunities. Some respondents expressed frustration with this instability, with a few opting to leave academia altogether. The study concludes that providing funding, and supporting network-building initiatives are critical for mitigating mobility's negative effects. Institutional policies fostering inclusiveness and career stability are essential to ensuring that international mobility benefits researchers across different disciplines.

Introduction and research relevance

The international mobility of highly productive researchers has emerged as one of the most widely discussed and debated topics within the global scientific community, primarily due to the multifaceted consequences of long-term mobility. Researchers have differing views on the advantages and disadvantages associated with long-term academic mobility. On one hand, long-term mobility is often associated with risks, career uncertainties, and the potential loss of vital social connections (Courtois & Sautier, 2022). On the other hand, it can offer significant opportunities for knowledge acquisition, enhance access to state-of-the-art equipment and new data,

as well as foster professional development and the expansion of research potential (Borini et al., 2018). The long-term mobility of researchers is thus a double-edged sword, offering both considerable benefits and noteworthy challenges.

Long-term mobility presents challenges for researchers in the social sciences and humanities. First and foremost, scholars in these fields often face the issue of lacking a universal language for international communication. Unlike in the natural sciences, where English predominates as the global language of communication, social scientists and humanists are often deeply connected to the symbolic and conceptual systems of the local communities where they conduct their research and publish their work. This disconnection from a global lingua franca presents significant barriers to cross-border academic exchange. For instance, the debate surrounding the use of national languages in academic publications is a point of contention in many countries. This issue is especially apparent in the tension between the use of Chinese and English. Despite a growing trend towards postcolonial discourse and efforts to "give voice to the oppressed" (Spivak, 2022), English remains "undoubtedly the preferred language in the social sciences and humanities" globally (Ammon, 2001, p. 10). The choice of publication language is influenced by a range of factors, including institutional constraints, established academic norms, the practices and ethics of the research community, and, importantly, the linguistic competencies of the researchers themselves (Canagarajah, 2002; Curry & Lillis, 2004). A further layer of complexity arises from the deeply contextual nature of social science and humanities research. The need to account for local cultural and stylistic nuances in language and adhere to the specific rules of the "language game" (Wittgenstein, 1985; Petersen & Shaw, 2002) creates a barrier for researchers attempting to disseminate their findings on an international scale. Failure to navigate these complex linguistic and cultural dynamics can undermine a scholar's professional credibility and hinder their career progression.

In addition to the challenges posed by language barriers, researchers in these fields also face the issue of cultural proximity, which imposes both formal and informal limitations on their ability to engage in international publication networks. A study of bibliographic networks among social scientists in Eastern Europe (Pajić, 2015) illustrates how national policy goals, such as integrating local research into international academic databases, drive the desire to publish in international journals. However, despite the increasing internationalization of communication channels, the processes of academic communication remain predominantly national and regional in nature. As a result, many Eastern European sociologists continue to rely heavily on national and regional journals for their publications, limiting their ability to engage with global academic networks. This trend creates a significant barrier to the globalization of research and hampers the integration of scholars in the humanities and social sciences into broader international research networks.

Furthermore, it is crucial to highlight that international research organizations tend to show greater interest from the natural sciences (Latova & Savinkov, 2012), while

those in the humanities and social sciences experience fewer tangible benefits from academic mobility. The discomfort many social scientists and humanists feel when encountering radically different approaches to disciplines such as history and sociology often diminishes the impact of international mobility on enhancing their research competencies or advancing their careers (Dyachenko & Nefedova, 2024). As a result, social science and humanities researchers often remain isolated within their local academic communities, thereby forming more insular networks that limit their engagement with professionals from other countries and regions. This situation presents additional barriers to scholars emigrating from Russia, as they are further distanced from global academic discourse.

The intellectual diversity within the social sciences and humanities, due to the creative and transformative nature of these fields, exacerbates this issue. According to the theory of scientific change, the lack of a unified research network contributes to intellectual and social fragmentation, with new data and innovative concepts being unevenly distributed across different regions (Fuchs, 1993). This fragmentation further complicates the career prospects of Russian scholars in the humanities, especially those who relocate abroad. In many cases, these researchers face significant challenges in securing relevant academic employment opportunities that align with their qualifications or professional standing. They are often offered positions that do not match their expertise or status, reflecting the limited recognition of humanities scholars on the international job market (Naumova, 2023).

The absence of universally recognized frameworks and symbols within the humanities and social sciences thus creates considerable obstacles for maintaining a successful academic career after emigration. For many scholars, gaining recognition in the global academic community is a more labor-intensive and challenging endeavor than it is for their colleagues in fields like engineering and natural sciences. For instance, Chinese scholars in the social sciences and humanities are far less visible in the international job market compared to their peers in the natural sciences (Flowerdew & Li, 2009).

Another significant challenge faced by researchers in the social sciences and humanities abroad is the increasing commercialization of academic fields. The growing focus on the potential for commercialization has profound implications for the career development of researchers in these disciplines, leading to several adverse consequences for both individual careers and the broader academic environment. This trend often exacerbates difficulties in securing research funding, with social scientists and humanists competing for limited resources within highly competitive institutional settings. The increased commercialization of academic work ultimately disrupts the academic climate, weakening scholarly connections and hindering collaborative efforts on joint projects (Leslie & Slaughter, 1997). This environment of intense competition, paired with a lack of sufficient funding and institutional support, can stifle the long-term growth and success of researchers in the social sciences and humanities.

Methodology

As a conceptual framework, this study employs the theory of three researcher careers (Gläser & Laudel, 2015). This theoretical model identifies three interconnected dimensions of career development that researchers navigate throughout their professional lives: the cognitive career, which pertains to their expertise, research competencies, and active engagement in scientific processes; the organizational career, encompassing their position, status, and career advancement within institutions; and the community career, which relates to their role and standing within the broader scientific community, including their professional networks and affiliations.

Despite its potential benefits, international mobility is not without challenges. A key issue lies in the lack of guarantees for long-term organizational stability. Temporary international assignments or fellowships often do not translate into permanent positions within research institutions, leaving scholars uncertain about their career trajectories. Moreover, during extended periods abroad, researchers may lose critical social connections within their home country's academic community. Upon returning, they often face the challenge of rebuilding their networks and readapting to local scientific environments. This readaptation process can weaken their standing within the community dimension, as they may struggle to reintegrate into professional networks and reestablish their influence. Consequently, these challenges often motivate researchers to seek further opportunities abroad, contributing to a brain drain phenomenon, where highly skilled individuals leave their home countries in search of more favorable conditions elsewhere.

To examine these dynamics, the study employs a qualitative research design, drawing on data collected between January 11 and May 3, 2024. The final dataset comprises 20 in-depth semi-structured interviews with researchers who met specific criteria. Participants were selected based on the following conditions: 1) Active involvement in research within the humanities and social sciences; 2) A history of long-term academic mobility, defined as sustained overseas academic engagement lasting more than one year. The qualitative approach allowed for a nuanced exploration of the interplay between the cognitive, organizational, and community dimensions in the context of international mobility. The interviews provided rich insights into how researchers navigate the complexities of career development, particularly the ways in which mobility influences their professional trajectories.

Preliminary Results

Due to the aforementioned challenges related to the mobility of social and humanities researchers, several scenarios of the outcome of mobility were reviled. Most of them, with rare exceptions, were related to losses in all three dimensions.

Rebuilding networks and professional identity

Because of the need to interact with the structures of everyday life and the social context, they feel the need to reconstruct their network of contacts and legitimise

their expertise in the new country. Reconstructing networks of contacts can be achieved through active engagement in social life in the new country. For example, one young researcher, who had many contacts thanks to conferences, admitted that his strategy was to attend various events that were not even indirectly related to his work:

"*And I thought that overall I already had some groundwork that could be realised, that could be useful to somebody here. On the other hand, during that time I got to know and understand some people in the local context, through whom I was able to settle down here. Although in the end it was not quite like that, because the local context is seen very differently from Russia, especially through the people who gave me access to the field, to the local context».*

Adaptation can take place in different ways, including atypical ways. For example, for one of the researchers, immersion theatre became a tool for understanding the social environment. It is interesting to note that this activity was not an attempt to compensate for stress or even an act of creative self-realisation; on the contrary, the respondent defined it as an "initial strategy":

"When I moved, I had a clear motivation to find new acquaintances. To be among people. For this purpose I chose from my activities in Moscow what seemed interesting, promising. I went to improvisation theatre. <...> I realised that in the confusion of the collective I would get the right feeling of life. The performances have a local texture, a local life. People talk about what is happening here and now. Very quickly you get a sense of context, a sense of where you are".

Shifting to low-skilled positions or precarious employment

The constant need to migrate in search of stable work takes a heavy emotional and professional toll on researchers. Many report feeling exhausted by the instability, leading some to accept less engaging or technical jobs to compensate for the negative effects of migration. One prominent anthropologist from Russia reflected on her decision to leave the field altogether:

"I have no energy left, my sociological curiosity is gone. I'm trying to find a more technical job that has nothing to do with academic work. This shift away from academic roles reflects the cumulative strain of navigating precarious employment conditions and the limited availability of suitable positions".

A recurring theme among respondents was the challenge of rebuilding professional networks in their host countries, a process that significantly diminished their standing within the academic community. Many noted that migration often resulted in the loss of professional privileges once enjoyed in their home country, where established connections facilitated access to resources and opportunities. One researcher lamented:

"*Yes, I have lowered my professional status, I have no administrative workload, no teaching, but I am still a sociologist. Now I do industrial sociology. Of course, I have significantly reduced my activity and my ability to do academic work. I've tried to write something, but the academic part of my life has come to nothing, I don't work*

on any clear-cut projects now. I mean, I used to have this grant, that grant, another grant, and this programme and that programme in parallel".

This loss of professional influence underlines the wider impact of migration on the community dimension of researchers' careers. The need to reestablish networks from scratch not only hinders career progression but also isolates researchers from key academic and professional ecosystems, further exacerbating the challenges of integration.

Moving to similar language contexts

Language emerged as a crucial factor influencing researchers' decisions about migration destinations. The ability to work and communicate in a familiar language often shaped their choices. For example, one respondent deliberately chose to migrate to Kazakhstan because of the opportunity to work in Russian: "*Weighing all the pros and cons, I finally chose Kazakhstan because it has the same educational programmes and I could work in Russian"*. For others, language skills and personal connections provided pathways to employment, albeit outside of research-intensive positions. One Sinologist reported securing a teaching position in Chinese through her network: "*I got a job teaching Chinese at a language school last autumn. I had to give up my research".*

Meanwhile, researchers who were unable to find positions directly related to their expertise turned to teaching Russian as a foreign language, a field in high demand on international labour markets: "*At the end of last year I realised that I couldn't find anything in my field <...>. At the local university, some courses were left unfilled due to a professor's maternity leave, and they gave me a course for this semester - it is Russian".*

Towards applicable science, neutral to the social context

Some of the researchers claimed that they wanted to change their specialisation to be more neutral to the reality of the social context. The most common scenario is to study some software to analyse data, for example:

"I'm upgrading my qualifications in some other related, even other fields, like data science. So what prevents me from feeling completely comfortable is the lack of universality in my professional activity. I realise that I need skills that would be useful absolutely everywhere, because so much is strangely specific, so I would like something more universal".

"In general, I see the biggest step in my situation is to learn Python and work as a data analyst. It seems like the most logical step. And the most important thing is that I will not find it uninteresting and I will acquire skills for myself".

Conclusion

Respondents described the positive aspects of mobility, such as opportunities to improve skills, exposure to different academic cultures and increased access to prestigious publication platforms. However, they also highlighted the significant challenges associated with their experiences. These included difficulties in maintaining long-term job security, the erosion of professional networks in their

home countries, and the emotional toll of adapting to new environments and academic cultures. The findings underline the dual nature of international mobility, highlighting both its potential to advance researchers' careers and its capacity to create significant barriers to long-term professional stability and integration.

The theory of three research careers provides a comprehensive framework for understanding the complex dynamics of career development in academia. The interplay between the cognitive, organisational and community dimensions highlights the multifaceted nature of researchers' careers, particularly in the context of international mobility. While mobility offers valuable opportunities for professional development and access to global resources, it also poses significant challenges, including career uncertainty, loss of social ties, and difficulties in reintegration. Addressing these challenges requires a deeper understanding of the unique experiences of mobile researchers and the development of institutional policies that support sustainable career development in all three dimensions.

Acknowledgments

The paper is based on the study funded by the Basic Research Program of the HSE University.

References

Ammon, U. (2001). The dominance of English as a language of science: Effects on other languages and language communities. Berlin: Mouton de Gruyter.

Borini, F. M., Lima, M. C., Pereira, R. M., & Siekierski, P. (2018). International academic mobility and innovation: A literature review. *Journal of Global Mobility*, 6(3-4), 285-298.

Borini, F. M., Oliveira, M. M. de, Bernardes, R. C., & Freitas, H. M. R. (2018). The impact of internationalization on the performance of emerging market multinational enterprises: The moderating role of innovation. *International Business Review*, 27(3), 542-554.

Canagarajah, A. S. (2002). A geopolitics of academic writing. Pittsburgh, PA: University of Pittsburgh Press.

Courtois, A., & Sautier, E. (2022). *Academic precarity: Understanding vulnerability, insecurity, and uncertainty in academia*. London: Routledge.

Curry, M. J., & Lillis, T. (2004). Multilingual scholars and the imperative to publish in English: Negotiating interests, demands, and rewards. *TESOL Quarterly*, 38(4), 663-688.

Dyachenko, A., & Nefedova, I. (2024). Academic mobility in the humanities: Challenges and consequences. *Russian Journal of Sociology and Anthropology*, 15(1), 34-49.

Flowerdew, J., & Li, Y. (2009). The globalization of scholarship: Major trends and emerging issues. *Journal of English for Academic Purposes*, 8(1), 44-57.

Fuchs, S. (1993). A sociological theory of scientific change. *Social Forces*, 71(4), 933-953.

Gläser, J., & Laudel, G. (2015). A heuristic framework for analyzing the careers of researchers. *Social Studies of Science*, 45(1), 89-110.

Latova, N. V., & Savinkov, V. I. (2012). Migration of scientists: International trends and Russian realities. *Sociological Research*, 51(5), 55-70.

Leslie, L. L., & Slaughter, S. (1997). The development and current status of market mechanisms in United States postsecondary education. *Higher Education Policy*, 10(3-4), 239-252.

Naumova, I. A. (2023). Academic emigration from Russia: Issues and prospects. *Russian Journal of Humanitarian Research*, 22(3), 78-88.
Nussbaum, M. C. (2019). *Not for profit: Why democracy needs the humanities*. Princeton, NJ: Princeton University Press.
Pajić, D. (2015). Globalization of the social sciences in Eastern Europe: Bibliometric perspective. *Scientometrics*, 102(3), 2131-2149.
Petersen, A. M., & Shaw, J. M. (2002). The cultural basis of scientific work: Contextual influences in comparative linguistics. *Linguistic Inquiry*, 33(2), 239-261.
Spivak, G. C. (2022). *Can the subaltern speak? Reflections on the history of an idea*. New York: Columbia University Press.
Wittgenstein, L. (1985). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Oxford: Blackwell. (Original work published 1953)

https://doi.org/10.51408/issi2025_125

How is the Sino-US AI Collaboration Reshaped by the China Initiative?

Dingkang Lin[1], Jiang Li[2]

[1] 652023140004@smail.nju.edu.cn, [2] lijiang@nju.edu.cn
Nanjing University, School of Information Management, 210023 Nanjing (China)

Abstract

The China Initiative, launched by the United States in 2018, has significantly reshaped scientific collaboration patterns between China and the U.S., particularly in the field of artificial intelligence (AI). This study examines the evolving dynamics of Sino-US AI research collaboration, focusing on the post-2018 period marked by geopolitical tensions, by using a comprehensive dataset from DBLP and DBLP-Citation-Network-v16. Our analysis reveals that (1) collaboration between the two nations shows a reversed U-shape where the peak is 2019, (2) China shifts its international collaboration to the EU and the U.S. strengthens ties with Canada, and (3) the AI subfield computer vision experiences the most pronounced impact under the China Initiative, because new collaboration in this field dramatically decreases and existing collaboration is largely suspended, which highlights its vulnerability to geopolitical disruptions.

Introduction

The Launch of the China Initiative in November 2018 in the United States (US) has significantly impacted scientific collaboration patterns of the US. US-based researchers have become increasingly cautious about engaging in collaborations with Chinese counterparts due to perceived risks and potential complications (Lee, 2022). This climate has particularly affected Chinese-American scientists, who have reported experiencing systemic discrimination and targeted scrutiny. The barriers to scientific mobility have become more pronounced, with Chinese scientists facing substantial obstacles in visiting US institutions. These challenges include increased visa denials and heightened bureaucratic hurdles, leading to a noticeable decline in their willingness to engage in US collaborations (Silver et al., 2020). Furthermore, the flow of Chinese students to US institutions has been significantly restricted, with limitations imposed on study fields and a marked decrease in enrollment numbers (Feder, 2019; Tang et al., 2021).

Prior to 2019, Sino-US research collaborations demonstrated consistent growth, predominantly funded by Chinese sources and characterized by a majority of Chinese first authorships (Lee & Haupt, 2020). However, post-2019 data reveals a concerning trend: both the absolute number of Sino-US collaborative publications and their proportion in global collaborative output have declined significantly (Tang, 2024; Wagner & Cai, 2022). At the individual researcher level, the China Initiative's impact is evident in productivity metrics. US scientists collaborating with China have experienced lower research output compared to those collaborating with other countries (Jia et al., 2024). Similarly, Chinese researchers engaged in US collaborations have shown decreased productivity and citation impact, prompting many to redirect their collaborative efforts toward domestic partnerships and

collaborations with other nations (Li & Wang, 2024). The differential impact of these collaboration shifts is particularly noteworthy given the varying degrees of reliance on international partnerships. US scientific innovation demonstrates greater dependence on international collaboration across multiple metrics, including patent filings and research publications (Jang et al., 2022; Wu et al., 2019). This suggests that the decline in Sino-US scientific cooperation may have more substantial implications for US research output and innovation capacity (Wagner & Cai, 2022). The field of AI (artificial intelligence) experienced increasing international collaboration prior to 2019, with the US and France maintaining central positions in global networks, while China emerged as hubs within developing countries' collaboration networks (Hu et al., 2020). However, recent trends indicate challenges. Okamura's (2023) global analysis observed declining multidisciplinary collaboration between China and the US post-2019, including in AI.

This study aims to systematically investigate the evolving dynamics of Sino-US collaboration in AI research, focusing on the question: *How has the China Initiative reshaped Sino-US collaboration in AI*? To address this question, we leverage a comprehensive dataset from DBLP. By categorizing AI research into ten distinct fields and employing robust methods for country attribution, we provide a nuanced analysis of collaboration trends, alternative collaborators of China/US, and potential explanations of the change, offering valuable insights into the broader implications for global AI innovation and scientific collaboration.

Methodology

Data Processing

The primary database used in this study is DBLP (Digital Bibliography & Library Project), an open-source bibliographic information database focused on major computer science publications. We retrieved the DBLP data on November 1, 2024. DBLP was chosen because it offers the most comprehensive collection of research papers published in both journals and conferences within the field of computer science.

The CCF (China Computer Federation) Recommended International Academic Publications Directory (2023 edition) lists 102 AI journals and conferences (CCF, 2023) (available at https://github.com/lindingkang/sino_us_ai_collaboration/blob/main/CCF_ai_conf_joun_2023.csv). Papers published in these venues were classified as AI research papers in our study. All journals and conferences but one are indexed in DBLP, i.e., Journal of Speech, Language, and Hearing Research. Consequently, our initial dataset includes 543,626 papers published in these 101 venues.

DBLP does not provide information on author affiliations. We hence utilized DBLP-Citation-Network-v16, developed by Tang et al. (2008), to augment our dataset in this regard. To address the inconsistencies in the writing of affiliations, we employed four distinct methods to determine the country of each author of the 1,388,182 author pairs from 440,797 articles that had complete records in the dataset: institutional

matching, country matching, manual matching, and AI-assisted matching, as follows,

- Institution-matching: Matching institutions to countries using OpenAlex institution information. We utilized the entire OpenAlex database, encompassing all institutions and their corresponding country information, to perform full-text matching with each author's affiliation texts using all available names, including those in different languages, alternative names, and other variants.
- Country matching: Using country names, aliases, and abbreviations for unmatched cases.
- Manual matching: Manually assigning countries to texts appearing over 30 times.
- AI-assisted matching: Inquiring with DeepSeek regarding the countries associated with the remaining affiliations.

As a result, a total of 1,165,155 texts were successfully matched, and upon conducting a manual verification of a 150-sample subset, we confirmed an accuracy rate of 100%. Following the mapping of all affiliations, we acquired 343,297 papers. By applying a filter for the years 2013 to 2022, we arrived at a final dataset comprising 180,821 articles authored by 237,741 individuals.

We constructed ten subfields of AI, by integrating the subfields from the AI Act by the European Union (https://artificialintelligenceact.com/understanding-ai-types-of-ai/) as well as insights from AI professionals, including machine learning, natural language processing, computer vision, cognitive computing, rule-based AI, robotics, multi-agent systems, expert systems, natural computing, and generative AI. Then, we categorized each journal or conference to one or more of the subfields according to the perspectives derived from large language models (LLMs) and AI professionals (available at https://github.com/lindingkang/sino_us_ai_collaboration/blob/main/CCF_ai_conf_joun_2023.csv). Papers published in a given journal or conference were assigned to the field(s) associated with that venue.

Measures

Okubo et al. (1992) proposed the *Affinity* index, defined as $C_{x,y}/Cx$, where $C_{x,y}$ represents collaborative publications between countries x and y, and C_x is country x's total international collaborations. In this study, we applied its variant to quantify the Sino-US collaboration, i.e.,

$$Affinity = \frac{C_{x,y}}{\sqrt{C_x * C_y}}. \quad (1)$$

We classified Sino-US author pairs in papers into two categories: *existing* collaboration and *new* collaboration. The delineation between old and new collaborations was anchored by the year 2019. Specifically, during the 2019-2022 period, an author pair in a paper was considered to have an *existing* collaboration if they had co-authored an AI-related paper in or before 2018. In contrast, if a pair had

no record of co-authoring any AI-related paper prior to 2018, it was labeled as *new* collaboration. During the 2013-2018 period, for a given year, an author pair in a paper was deemed to have an existing collaboration if they had previously co-authored an AI-related paper before that year. Conversely, if there was no prior record of them co-authoring any AI-related paper before the given year, the pair was classified as a new collaboration.

Based on the definitions above, we classified papers into two groups. Papers where all author pairs were *existing* collaborations were labeled as *existing* collaborations, while papers that included at least one *new* author pair were classified as new collaborations. In the computation of the affinity index, the denominator remained the total number of international collaborative publications between the two countries, while the numerator was the count of either new or existing collaborative papers between them. It should be emphasized that the combined count of *new* and *existing* pairs is not equivalent to the overall number of author pairs from 2019 onwards. This discrepancy signifies author pairs that initially emerged post-2018 but then reoccurred in later years, thereby illustrating the evolving characteristics of collaborative relationships across different time periods.

Furthermore, we characterized *disappeared* collaboration as referring to Sino-US author pairs who were present in papers published in or before 2018 but were absent from publications in any year subsequent to 2018.

Result

China has seen a more pronounced increase in the quantity of AI publications compared to the US, as illustrated in Figure 1. Although the number of US AI publications rose sharply, a downturn emerged in 2022. In terms of international collaboration on AI publications, both China and the US have experienced growth, but this growth rate is slower than that of the overall number of AI publications.

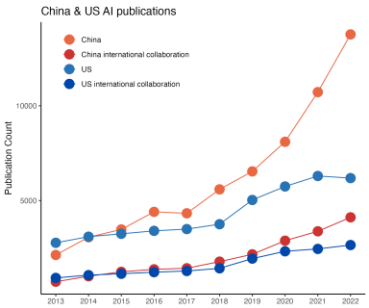


Figure 1. Trends in AI publications of China and US between 2013 and 2022.

Following the initiation of the China Initiative in 2018, there has been a significant rise in Sino-US collaboration in terms of AI publication counts, despite a slight deceleration in the growth rate, as depicted in Figure 2a. In parallel, Sino-EU collaboration in AI publication counts has been on a steady upward trajectory. Regarding the US, its collaborations with both the EU and Canada have increased, albeit at a pace that is not as rapid as that observed with China. To a certain extent, the upward trend in all four of these collaboration curves in Figure 2a can be

attributed to the overall increase in the number of AI publications in China and US, as shown in Figure 1. Accordingly, we leverage the Affinity index to mitigate the effect of publication counts.

The upward trajectory of Sino-US collaboration in AI (measured by the Affinity index) shifted dramatically to a downward trend, forming a reversed U-shape as illustrated in Figure 2b. The peak in the curve in 2019 may likely be due to publication delays. In contrast, China redirected its international collaboration towards the EU among all other countries/territories, while the US shifted its focus to Canada. The EU is not an alternative collaborator for the US, as their collaboration has seen a significant decline since 2019.

Next, we turn our attention to the declining Sino-US collaboration. Figure 2c demonstrates that both *existing* Sino-US collaborations and *new* collaborations (measured by Sino-US author pairs) have been sharply decreasing since 2019. It is important to note that the values in Figure 2c do not represent the number of Sino-US co-authored publications, but rather the Affinity index of Sino-US collaboration, which is divided into *new* and *existing* categories. Clearly, the decline in *existing* collaborations is more pronounced than that of *new* collaborations, suggesting that the overall decrease in Sino-US collaboration is primarily due to the contraction of *existing* collaborative relationships.

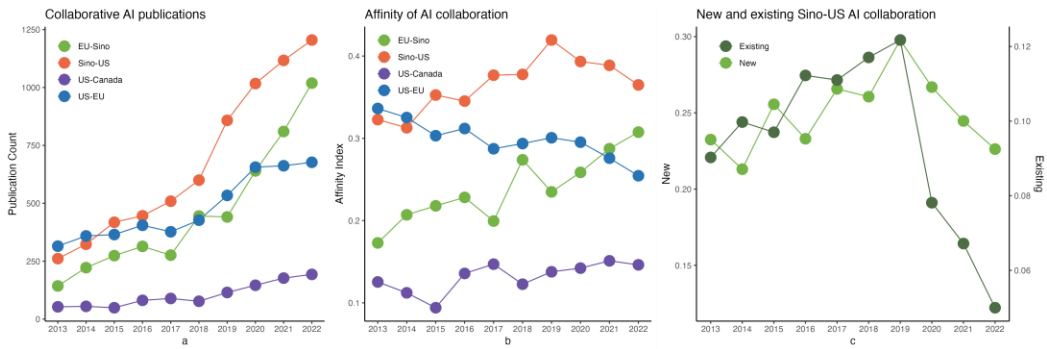


Figure 2. Sino-US collaboration in AI publications. 2a. Trends of Sino-US collaborative AI publications. 2b. Trends of Sino-US AI collaboration measured by the Affinity and their alternative collaborators. 2c. Trends of *new* and *existing* Sino-US AI collaboration.

Upon examining the subfields of AI, we observe that *new* collaborations in the majority of these subfields follow the general trend, with a steady decline since 2019. The most significant decreases are seen in robotics, natural language processing, and cognitive computing, with respective decline rates of 40.1%, 35.9%, and 35.9%, as depicted in Figure 3a. Figure 3b indicates that *existing* collaborations in most subfields also underwent a rapid decrease from 2019 to 2022, with rule-based AI, computer vision, and expert systems being the most impacted, experiencing decline rates of 61.9%, 61.6%, and 57.4% respectively.

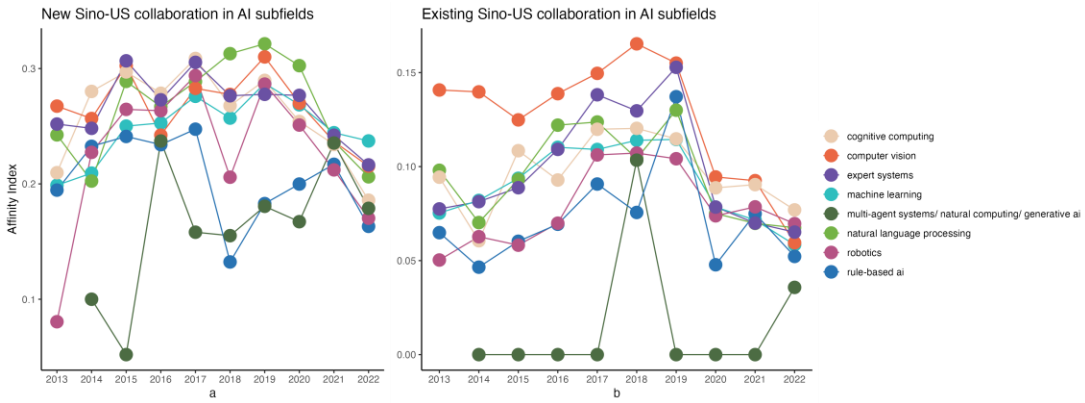


Figure 3. Trends of Sino-US collaboration in ten AI subfields (Due to limited publications, the fields of "multi-agent systems," "natural computing," and "generative AI" are combined into one). 3a. Affinity index of *new* collaboration in each AI subfield. 3b. Affinity index of *existing* collaboration in each AI subfield.

The decline of existing collaborations in Figure 2c indicates the disappearance of Sino-US collaboration. It is verified that a significant number of Sino-US author pairs have vanished since 2019, as indicated in Figure 4a. The count of such *disappeared* pairs skyrocketed to over 5,000 in 2020 and has stayed at a high level since then. Upon examining the subfields, it was found that computer vision was the most heavily impacted area, as illustrated in Figure 4b.

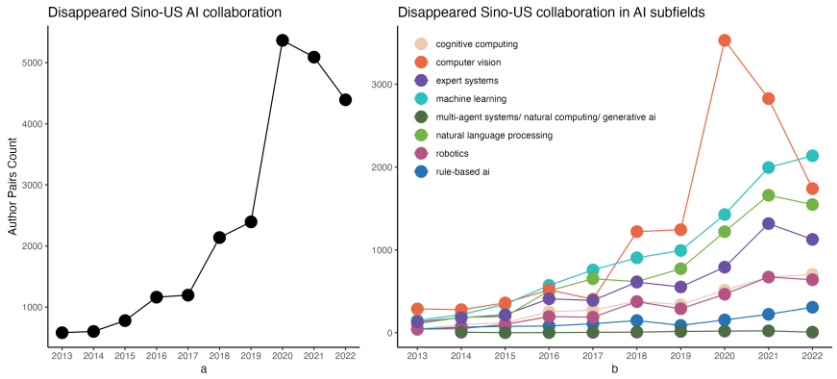


Figure 4. Trends of disappeared Sino-US AI collaboration. 4a. Disappeared Sino-US AI collaboration. 4b. Disappeared Sino-US collaboration is AI subfields.

Preliminary findings

This study provides a comprehensive analysis of the Sino-US collaboration in AI research after the launch of the China Initiative in 2018, leveraging 101 AI-related journals and conferences indexed in DBLP and DBLP-Citation-Network-v16. We delved into the ten distinct AI subfields to explore why changes happened.

The initial findings are as follows: (1) we identified a reversed U-shaped pattern in Sino-US AI collaboration from 2013 to 2022, with the peak occurring in 2019. The significant decline in Sino-US collaboration can be attributed to a sharp reduction in

both *new* and *existing* collaborative efforts. (2) In response to the China Initiative, China has turned to the EU as an alternative partner in AI, while the US has primarily looked to Canada for collaboration. (3) The AI subfields of computer vision has been most heavily affected by the China Initiative. This is due to a steep decrease of *new* collaborations and a near suspension of *existing* collaborations.

This study offers initial statistical insights, with the analysis grounded in observational findings rather than causal inferences. Moving forward, we aim to apply a difference-in-differences approach to rigorously establish causality, validate the current observations, and delve deeper into the underlying factors driving these trends.

Acknowledgments

This study is financially supported by the National Social Science Fund Major Project (24&ZD321). We extend our sincere gratitude to Dr. Yanbo Wang for his constructive comments, and Professor Jie Tang and his team for their invaluable assistance with data access and processing.

References

CCF. (2023). *CCF Recommended International Academic Publications Directory*. https://www.ccf.org.cn/Academic_Evaluation/By_category/

Feder, T. (2019). Trade wars and other geopolitical tensions strain US-China scientific collaborations. *Physics Today, 72*(11), 22-26. https://doi.org/10.1063/Pt.3.4338

Hu, H. T., Wang, D. B., & Deng, S. H. (2020). Global Collaboration in Artificial Intelligence: Bibliometrics and Network Analysis from 1985 to 2019. *Journal of Data and Information Science, 5*(4), 86-115. https://doi.org/10.2478/jdis-2020-0027

Jang, B., Choung, J. Y., & Kang, I. (2022). Knowledge production patterns of China and the US: quantum technology. *Scientometrics, 127*(10), 5691-5719. https://doi.org/10.1007/s11192-022-04478-4

Jia, R. X., Roberts, M. E., Wang, Y., & Yang, E. D. (2024). The impact of US-China tensions on US science: Evidence from the NIH investigations. *Proceedings of the National Academy of Sciences of the United States of America, 121*(19). https://doi.org/10.1073/pnas.2301436121

Lee, J. J., & Haupt, J. P. (2020). Winners and losers in US-China scientific research collaborations. *Higher Education, 80*(1), 57-74. https://doi.org/10.1007/s10734-019-00464-7

Li, M. X., & Wang, Y. (2024). Influence of political tensions on scientific productivity, citation impact, and knowledge combinations. *Scientometrics, 129*(4), 2337-2370. https://doi.org/10.1007/s11192-024-04973-w

Okubo, Y., Miquel, J. F., Frigoletto, L., & Dore, J. C. (1992). Structure of International Collaboration in Science - Typology of Countries through Multivariate Techniques Using a Link Indicator. *Scientometrics, 25*(2), 321-351. https://doi.org/Doi 10.1007/Bf02028090

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). *ArnetMiner: extraction and mining of academic social networks* Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA. https://doi.org/10.1145/1401890.1402008

Tang, L. (2024). Halt the ongoing decoupling and reboot US-China scientific collaboration. *Journal of Informetrics, 18*(2). https://doi.org/10.1016/j.joi.2024.101521

Tang, L., Cao, C., Wang, Z., & Zhou, Z. (2021). Decoupling in science and education: A collateral damage beyond deteriorating US-China relations. *Science and Public Policy, 48*(5), 630-634. https://doi.org/10.1093/scipol/scab035

Wagner, C. S., & Cai, X. (2022). Changes in co-publication patterns among China, the European Union (28) and the United States of America, 2016-2021. *arxiv*. https://doi.org/https://arxiv.org/abs/2202.00453

Wu, L. F., Zhu, H. Y., Chen, H. C., & Roco, M. C. (2019). Comparing nanotechnology landscapes in the US and China: a patent analysis perspective. *Journal of Nanoparticle Research, 21*(8). https://doi.org/10.1007/s11051-019-4608-0

https://doi.org/10.51408/issi2025_110

How systematic are the systematic reviews?

Andrey Guskov[1], Denis Kosyakov[2], Irina Selivanova[3], Alexandra Malysheva[4]

[1]*guskov.andrey@gmail.com*
Russian Centre for Scientific Information, Leninsky pr., 32A, Moscow (Russia)
Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
Ac. Lavrentieva ave. 6, Novosibirsk (Russia)
Russian Institute of Economics, Policy and Law, Dobrolubova Str. 20A, Moscow (Russia)

[2]*kosyakov@sciencepulse.com*
Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
Ac. Lavrentieva ave. 6, Novosibirsk (Russia)
Russian Institute of Economics, Policy and Law, Dobrolubova Str. 20A, Moscow (Russia)

[3]*i-seli@yandex.ru*, [4]*bag_bala@mail.ru*
Russian Institute of Economics, Policy and Law, Dobrolubova Str. 20A, Moscow (Russia)

Abstract

Systematic literature reviews (SLRs) are widely recognized as a cornerstone of evidence-based research, providing comprehensive syntheses of existing literature on specific topics. Despite the availability of standardized protocols (e.g., PRISMA), many authors do not fully adhere to established methodological requirements. This study aims to determine how frequently four basic criteria – explicit search strategies, inclusion/exclusion criteria, a complete list of included sources, and a clear model of analysis – are met in publications that are labeled as SLRs.

Using Scopus, we sampled 1000 publications in four disciplines (Medicine, Computer Science, Social Sciences, and Biochemistry) and used large language models to assess compliance with each criterion. Results show that 53% of SLRs satisfy all four requirements, while 16% fail at least two. Search and inclusion criteria are widely recognized as core components of SLRs, while fewer authors provide a complete reference list or adopt an explicit analysis model. Disciplinary differences emerged, with Biochemistry and Medicine having the highest rates of full compliance, and Computer Science the lowest. In Medicine, high-impact journals had a 13% higher compliance rate, demonstrating the impact of journal policies. However, overall compliance did not correlate with citation impact. The prevalence of PRISMA in Medicine and Biochemistry likely drives higher compliance in these fields. Future research will expand the analysis by incorporating additional criteria and expert assessments, providing deeper insight into the role of SLR methodologies and the accuracy of evaluations based on AI-tools.

Introduction

Systematic literature reviews are considered to be one of the main tools of scientific methodology, as they summarize and critically analyze all available literature on a particular topic, forming a reliable evidence base for further research (Mathew, 2022). One of the most important principles of SLR is considered to be comprehensive sourcing, which promotes unbiased conclusions and reduces the risk of missing relevant data (Cooper et al., 2018), which can lead to biased effect estimates and unreliable conclusions (Tricco et al., 2008).

Despite the importance of methodological rigor and the availability of the well-known PRISMA family of protocols, many authors do not always adhere to these

requirements. For example, Norling et al (2023) showed that a large proportion of urology reviews did not report detailed search strategies. A further problem is the lack of detail in the description of inclusion and exclusion criteria: although authors often mention such criteria, the actual details of their application remain unclear (Budgen et al., 2018). Frost et al. (2022) also found that only 8% of protocols met all PRISMA-P requirements, indicating the formal nature of adherence to established methodological standards. Finally, many reviews ignore the recommendation to publish a full list of included sources (Kitchenham et al., 2022) and limit themselves to a general description. As a result, it is not uncommon for reviews that claim to be 'systematic' to actually have a very superficial methodology, while some 'mapping studies' are closer to full-fledged SLRs (Budgen et al., 2018).

Large Language Models (LLMs) are increasingly being used to process the growing amount of scientific information. There are already examples of their successful use to automate the processes of selection, extraction, judgment, analysis and narration in the preparation of SLR, which show results comparable to those of experts (Hasan et al., 2024). However, it remains an open question to what extent review authors themselves correctly specify and apply the underlying methodological principles when assessing the quality of such reviews against the key criteria of transparency and reproducibility. In particular, Budgen et al. (2018) showed that review authors do not always fully and transparently describe the sourcing, inclusion/exclusion, list of selected primary studies, and data analysis model, even though these aspects directly affect the reproducibility of reviews and provide a basis for assessing their methodological quality. However, systematic peer review of these requirements is laborious, making it difficult to regularly analyze the quality of SLRs.

The **aim** of the study is to test, using large language models, how often basic requirements are met in SLRs that are labeled as systematic:

- **R1:** presence of explicitly stated criteria for finding sources,
- **R2:** presence of explicitly stated criteria for inclusion/exclusion of sources,
- **R3:** presence of a list of sources selected for review,
- **R4:** presence of a model for the analysis of sources.

Based on this objective, the **following research questions** are formulated:

RQ1. For what part of SLRs are requirements R1-R4 fulfilled?

RQ2. Are there statistically significant differences in compliance between disciplines?

RQ3. Are these requirements more often fulfilled in high-impact journals?

RQ4. Is there a relationship between completing requirements and citing SLRs?

Method

Four scientific fields were selected for the study in which SLRs have a significant representation (ASJC code in parentheses):

- **Medicine** (2700) has the longest tradition of standardized systematic reviews, particularly under the PRISMA guidelines, and exhibits clear protocols for risk of bias assessment and data synthesis.
- **Computer Science** (1700) has experienced a rapid increase in the number of SLRs, often adapting methodologies from other fields or employing

alternative frameworks such as Kitchenham's guidelines, thus illustrating a discipline in the midst of methodological standardization.

- **Social Sciences** (3300) represent a broad, interdisciplinary arena where systematic reviews are also undertaken but are typically governed by more flexible or mixed-method approaches, providing a contrast to the highly codified medical SLR protocols.
- **Biochemistry** (1300) typifies a natural science discipline that frequently employs SLRs to summarize experimental evidence; it also increasingly intersects with data-driven analyses, making it pertinent for assessing how LLMs handle specialized literature.

In each of these areas, a sample of publications was generated from Scopus that met the following criteria:

- Title or abstract contains "systematic review" OR "systematic literature review",
- Publication year 2022,
- Document type 'Article', 'Review', or 'Conference Paper',
- Open access (any).

From each sample, 400 publications were randomly selected. These were pre-filtered using LLM gpt-o1-mini: the title and abstract were checked to ensure that they were indeed systematic reviews in the specified scientific field. For those that passed, the full text of the publications was downloaded. The text layer was extracted from the PDFs and the number of tokens was calculated (model cl100k of the Python library tiktoken). Publications that appeared to have less than 2,000 or more than 50,000 tokens were discarded. From the publications that passed all checks, 250 were randomly selected for each discipline and a final sample (N=1000) was drawn.

For each article in this sample, the gpt-4o language model was used to determine whether R1-R4 requirements were met, as well as mentions of SLR preparation techniques. Sampling of the results by the article authors showed a satisfactory result.

Results

RQ1. For what part of systematic reviews are requirements R1-R4 fulfilled?

All four requirements are met in 53% of the reviews and 16% of the reviews, which the authors call systematic, do not meet 2 or more requirements (Figure 1). The requirements to specify criteria for finding publications (R1, 89%) and to include them in the review (R2, 93%) are most frequently fulfilled. This is not surprising, since in many journals the requirement to specify where and how publications were searched for and according to which principles they were selected has already become the "gold standard" for SLRs, regardless of the field.

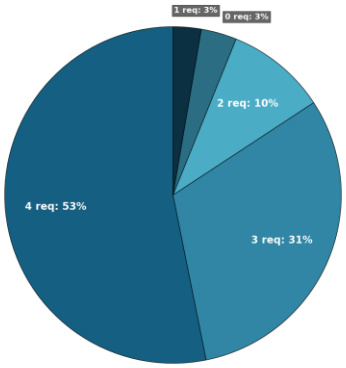


Figure 1. Distribution of SRLs by number of requirements met.

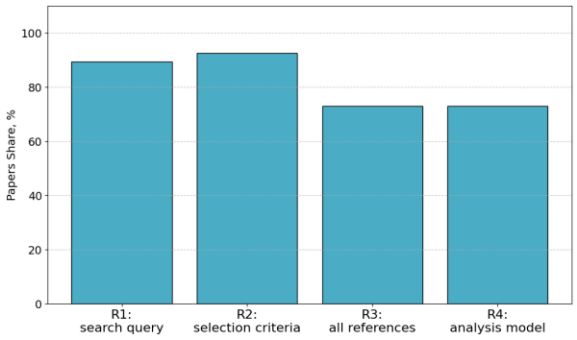


Figure 2. Degree of fulfillment of requirements, entire sample.

The other two requirements 'all references' and 'analysis model' are less frequently fulfilled - only in 73% of the cases each (Figure 2). Moreover, if we consider only 13% of the publications that did not meet exactly two requirements, the majority of them fall, as expected, on this pair (53 out of 93 cases). Most likely, such simplifications are made by authors who do not bother to formalize the analysis and do not see the need to provide an exact list of included articles. This practice is more typical in more "liberal" or interdisciplinary fields, or where journals do not impose strict requirements.

Failure to comply with two or more requirements may also indicate a lack of awareness of common standards among authors and the absence of rigid review filters in relevant journals and conferences.

RQ2. Are there statistically significant differences in compliance between disciplines?

When analyzing the fulfillment of the requirements for SLRs in different disciplines, certain differences can be observed (Figure 3). For example, all four requirements are most often met in biochemistry (65%) and medicine (63%), and least often in computer science (38%). Conversely, in the first two fields it is extremely rare not to meet any of the requirements (1%), while in computer science it is not so rare anymore (10%). It should be noted that in this field, each requirement is fulfilled much less frequently than in the other fields.

This difference can be explained by the fact that the medical sciences have already established a "gold standard" – the PRISMA family of protocols – which prescribes these and other requirements for SLRs. Our study showed that in biochemistry and medicine,

80-85% of reviewed publications follow these protocols. It is so widespread that it has already penetrated deeply into many disciplines, including the social sciences (64%) and computer science (55%).

In the latter, an alternative methodology known as Kitchenham's guidelines (Kitchenham & Charters, 2007) is sometimes encountered (4%). Mentions of other methodologies occurred 1-2 times (totaling 1.5% of the sample) and were not

included in the analysis. Overall, this suggests that adherence to review across disciplines is related to the prevalence of the PRISMA standard.

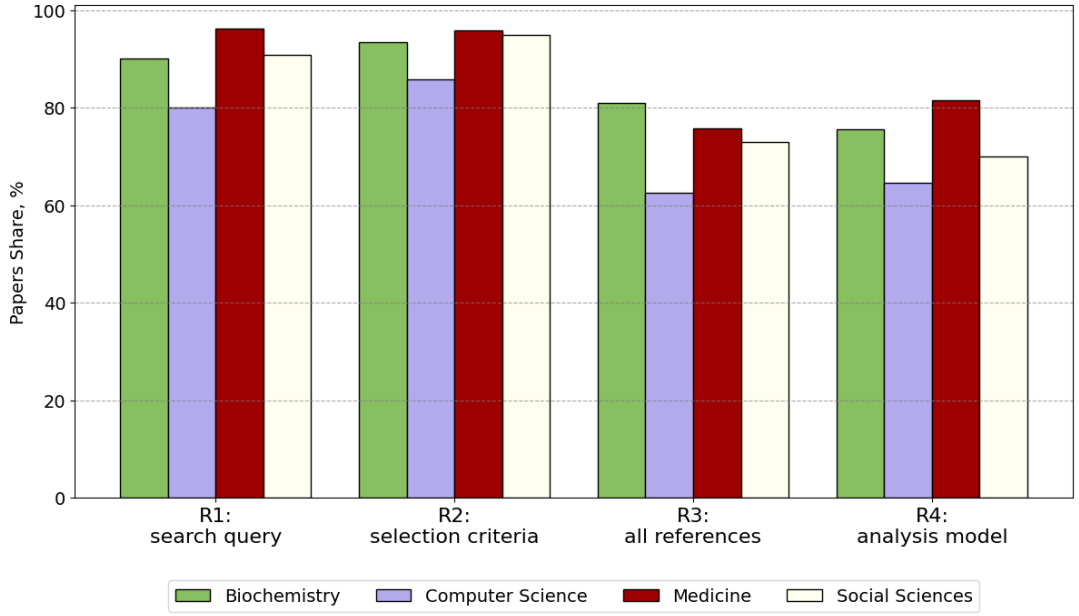


Figure 3. Degree of fulfillment of requirements by field of science

RQ3. Are these requirements more often fulfilled in high-impact journals?

At this stage of the study, high-impact journals are considered to be those that are in the top 10% of journals in a given scientific field according to the SJR (SCImago Journal Rank). Publications in other journals were used as a control group (Other). For each group, the proportion of publications that met all four requirements was calculated.

As can be seen in Figure 4, only in medicine were there significant differences: the high-impact journals met all requirements 13% more than the other journals (73% vs. 60%). This suggests that checking compliance with the requirements considered here (more precisely, the PRISMA requirements) is part of the editorial policy of leading medical journals.

In the other three areas, the difference is less than 2% – it is likely that the practice of strict adherence to systematic review methodologies has not yet taken hold in these areas, as journal editors do not prioritize it. In this case, an interesting phenomenon can be observed in biochemistry, where the PRISMA standard is recognized by the scientific community, but the editorial policies of leading journals are not affected. Another explanation could be that in these disciplines' other approaches (e.g. "mapping studies" or "narrative reviews") are used to prepare SRLs and therefore there is no need to insist on strict compliance with all formal criteria.

Thus, in medicine, high-impact journals play a more stringent "regulatory" role, ensuring that SLRs meet all criteria for methodological transparency.

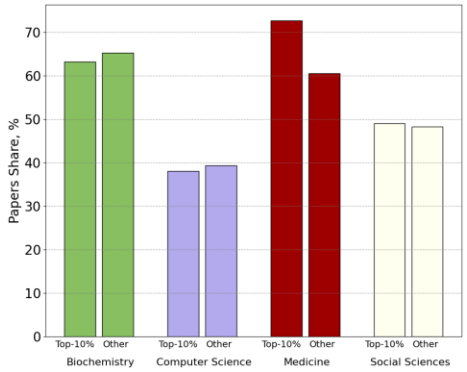


Figure 4. Percentage of SRLs that have all requirements met; comparing the 10% with the highest SJR to the others.

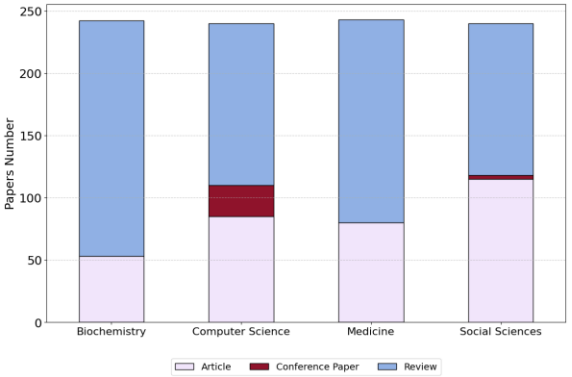


Figure 5. Distribution of SRLs by document type.

RQ4. Is there a relationship between completing requirements and citing SLRs?

After dividing the articles into groups according to the number of requirements fulfilled in them, we analyzed the distribution of field-weighted citation impacts obtained as of 2025/01/10. Contrary to our expectations, we found no significant differences in these groups, nor in those organized by discipline. It should be noted that the publications in question had a short "life cycle", whereas methodologically "high quality" papers may be recognized with a delay. However, we believe that methodological quality has little influence on the citation of the SLR, while more important factors are the relevance of the topic, the novelty or even the "brilliance" of the results, as well as the network of scientific communication and the authority of the authors. In the computer or social sciences, conceptual novelty, interdisciplinary scope, or practical implications may be more important than strict adherence to methodological guidelines. This is not to say that such reviews are not useful, but rather to distinguish between the notions of quality, relevance, and methodological rigor.

We feel it is necessary to highlight another important result. When searching for reviews in bibliographic review databases, a faceted filter by document type (doc_type=Review) is often used. Figure 5 shows that this results in filtering out 20 to 50% of publications that are also reviews, but of type Article. In addition, it is common practice in computer science to publish SLRs in conference proceedings with corresponding document types. There are also opposite situations where a document of type Review is not such a document. All this speaks not only about the imperfection of the mechanism of assigning document types in Scopus, but also about the mixing of two aspects in one doc_type attribute: source type (article for journals, CP for conferences, chapter for books) and content type (review, conference review, short survey, report). A complete solution to this problem is probably to separate these aspects into two different attributes and to clarify the rules for filling them in. Under the current conditions, we recommend not to filter by document type when systematically searching for reviews in Scopus.

Conclusion

Despite the widespread use of the PRISMA family of protocols, in practice there is still a certain "dis-synchronization" between what authors declare to be a "systematic review" and what is actually implied in the methodological guidelines. At the same time, the vast majority of authors already consider the search and inclusion criteria (R1, R2) as mandatory components of a SLR. However, a more detailed adherence to formal standards is not always realized, especially in fields with a less formalized methodological culture.

The presented results are preliminary. In the next phase of the study, we plan to expand the set of requirements under examination and to explore how their fulfilment relates both to the review methodologies employed and to the scope of the reference lists. The comprehensive list of requirements may eventually encompass all elements outlined in PRISMA – especially since *Frost (2022)* provides expert evaluation guidelines that could be adapted as prompts for LLMs. However, it should be noted that at present LLMs may not yet be able to thoroughly review all possible requirements, so the final set of criteria will need to be refined. A representative sample of SLRs will be peer reviewed using a similar methodology and the consistency of their results with the LLM data will be analyzed. As a result, the statistical significance of the results will be assessed. The project materials will be made available on GitHub.

References

Budgen, D., Brereton, P., Drummond, S. & Williams, N. (2018). Reporting systematic reviews: Some lessons from a tertiary study. *Information and Software Technology*, 95, 62–74.

Cooper, C., Booth, A., Varley-Campbell, J. *et al*. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. *BMC Med Res Methodol* 18, 85 (2018). https://doi.org/10.1186/s12874-018-0545-3

Frost, A. D., Hróbjartsson, A., & Nejstgaard, C. H. (2022). Adherence to the PRISMA-P 2015 reporting guideline was inadequate in systematic review protocols. *Journal of Clinical Epidemiology*, 150, 179–187. https://doi.org/10.1016/j.jclinepi.2022.07.002

Hasan, B., Saadi, S., Rajjoub, N. *et al* (2024) Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment *BMJ Evidence-Based Medicine* 2024;29:394-398. https://doi.org/10.1136/bmjebm-2023-112597

Kitchenham, B. & Charters, S. (2007) Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Kitchenham, B., Madeyski, L., & Budgen, D. (2023). SEGRESS: Software Engineering Guidelines for REporting Secondary Studies. *IEEE Transactions on Software Engineering*, 49(3), 1273-1298.

Mathew, J.L. (2022) Systematic Reviews and Meta-Analysis: A Guide for Beginners. *Indian Pediatr* **59**, 320–330. https://doi.org/10.1007/s13312-022-2500-y

Norling B, Edgerton Z, Bakker C, Dahm P. (2021) The Quality of Literature Search Reporting in Systematic Reviews Published in the Urological Literature (1998-2021). *Journal of Urology*, 209(5), 837-843. https://doi.org/10.1097/JU.0000000000003190.

Tricco, A. C., Tetzlaff, J., Sampson, M. et al. (2008). Few systematic reviews exist documenting the extent of bias: A systematic review. *Journal of Clinical Epidemiology*, 61(5), 422–434.

https://doi.org/10.51408/issi2025_148

Impact of Marriage on Productivity and Career of Women Scholars

Shiqi Tang[1], Xianjiang Deng[2], Jianhua Hou[3], Cassidy R. Sugimoto[4]

[1]*stang356@gatech.edu*, [4]*sugimoto@gatech.edu*
Georgia Institute of Technology, School of Public Policy, 258 4th Street Atlanta, GA 30332 - 0345 (United States)

[2]*dengxj27@mail2.sysu.edu.cn*, [3]*houjh5@mail.sysu.edu.cn*
Sun Yat-sen University, School of Information Management, No. 132, Outer Ring East Road, Guangzhou (China)

Abstract

Marriage has potential impact on scholarship, especially for women, but lack of appropriate data has prevented its clear assessment. In this article we quantify the impact of marriage on women's scholarship using open data from ORCID (23057 married women scholars are recognized), including longitudinal productivity data and career path. So far we have find marriage have short term negative impact but long term active impact on productivity of women scholars and this impact varies according to the field they worked in. The short term negative impact is more significant if they get married after starting their careers. While we continue to investigate other aspects of this topic, such as the impact of marriage on career progression, we believe this research will offer valuable insights for academic institutions and policymakers, helping to ensure that marriage does not become an insurmountable barrier to women's academic success.

Introduction

Marriage can significantly influence career trajectories, and its impact on women scholars is particularly worth investigating due to the unique demands of academic work (Juraqulova, Byington et al. 2015). The long and nonlinear career progression, reliance on research productivity for tenure and promotion, and the expectation of geographic mobility for academic appointments can create additional challenges for women balancing family responsibilities (Mantai and Marrone 2023). While previous studies have investigated the effects of marriage and parenthood on academic careers, they have primarily relied on survey data, with relatively few studies leveraging large-scale datasets for quantitative analysis. Establishing the causal impact of marriage on productivity and career progression has been challenging due to the lack of detailed longitudinal data on marriage timing, research productivity, and career transitions.

Marriage can lead to many changes to careers. On the one hand, marriages creates a new demand to allocate time to family, particularly for women, to support their family in housework or take care of children (Mason and et al. 2004, Schiebinger and et al. 2010). Parenthood has been proven to decreases the available research time for women, leading to drop down in their productivity(Joecks, Pull et al. 2013, Lutter and Schröder 2019). Parenthood has been shown to reduce research time and lower productivity, but the direct impact of marriage—independent of parenthood—

remains underexplored. Moreover, marriage may influence career trajectories beyond productivity. Women in academia may experience structural and cultural barriers that make academic careers less accommodating after marriage, leading to self-selection out of academia or shifts in job roles to better balance family responsibilities (Hawks and Spade 1998, Wolfinger and Goulden 2008, Cech and Blair-Loy 2019). Aslo considering the limited availability of faculty positions, it can be challenging for both spouses to secure academic jobs in the same city, which is also a factor driving some women scholars to transition from academia to industry.

This study aims to fill this gap by systematically quantifying the impact of marriage on women scholars' productivity and career progression. Using large-scale longitudinal data, we analyze: (1) the impact of marriage on research productivity, (2) the effect of marriage on career promotions, (3) how marriage influences career transitions between academia and industry, and (4) the evolving trends in these impacts over time. We investigate on the productivity pattern before and after marriage as an intervention event occur and compare the carreer trajectories of married women faculty with a selected control group of women faculty to investigate on the impact of marriage on career promotions and transitions. By employing rigorous causal inference methods, we provide a comprehensive analysis of how marriage shapes women's careers in academia and beyond.

Method

In obtaining data on married women researchers, we use ORCID open data to extract marriage timing and longitudinal productivity data, as well as career paths of female researchers. The identification of whether a female researcher is married and the timing of marriage is based on the following measure: In some countries and regions, female researchers change their surname to their husband's surname after marriage. ORCID records each user's name and name change history. We first identify female researchers based on their first names. If their surname undergoes a reasonable change (e.g., replacing their original surname with their husband's surname or adopting their husband's surname while keeping their original surname as a middle name), we consider this as an indicator of marriage. After data preprocessing, 23,057 married women scholars are identified.

We use the Regression Discontinuity Design (RDD) method to investigate the impact of marriage on productivity while controlling for individual fixed effects. RDD determines causal effects by assigning a cutoff or threshold above or below which an intervention is applied. Here, we consider marriage as the intervention, the annual publications as representation of productivity while assuming that there would be one year delay for the effect of marriage on productivity, since article publication needs time.

To analyze the impact of marriage on career promotion and transitions between academia and industry, further causal inference requires constructing an appropriate control group to match with the married women researchers. Therefore, we use Coarsened Exact Matching (CEM). The attributes used for matching include field, academic age, annual publication patterns at different academic ages, and career stage. Academic age is measured by the time elapsed since the first publication.

Result

Impact of marriage on productivity

Overall, marriage is associated with a reduction in the productivity of female scholars. Specifically, the impact varies across different fields: Technology (coefficient = -0.057), Physical Sciences (coefficient = -0.265*), Social Sciences (coefficient = -0.264*), Life Sciences & Biomedicine (coefficient = -0.220**), and Arts & Humanities (coefficient = -0.520**) as showed in Table 1. When examining different career stages, the impact of marriage is negative for women scholars get married during the work phase (coefficient = -0.240***), which is showed in Table 2.

Table 1. Impact of marriage on annual publications in various fields[1].

Field	*Publication*	*Coefficient*
Total	lwald	-0.053
	lwald50	(omitted)
	lwald200	-0.214***
Arts & Humanities	lwald	-0.520**
	lwald50	(omitted)
	lwald200	-0.305*
Life science & Biomedicine	lwald	-0.220**
	lwald50	-0.064
	lwald200	-0.231***
Physical Science	lwald	-0.265*
	lwald50	0.074
	lwald200	-0.133
Social science	lwald	-0.264*
	lwald50	-0.112
	lwald200	-0.240***
Technology	lwald	-0.057
	lwald50	(omitted)
	lwald200	-0.028

Table 2. Impact of marriage on annual publications during various time periods[1].

Period	*Publication*	*Coefficient*
Education	lwald	-0.231
	lwald50	-0.003
	lwald200	-0.100*
Employment	lwald	-0.240***
	lwald50	-0.057
	lwald200	-0.229***

Impact of marriage on career promotion

Married researchers take an average of 2 months longer to be promoted to associate professor and 6 months longer to be promoted to professor.

[1] Result of Regression Discontinuity Design. Lwald, lwald50, and lwald200 represents results in optimal estimating bandwidth, half of optimal estimating bandwidth, and double of optimal estimating bandwidth.

Trends in the impact of marriage

Figure 1 shows the cumulative number of publications over time relative to the first publication for two groups: "married" and "randomly selected". The trend lines suggest that, in most fields, there would be a specific time period that married women scholars publish less than their random selected counterpart, showing a potential negative effect of marriage. However, in some fields like Physical Sciences and Life Sciences & Biomedicine, married scholars continues to publish slightly more compared to their randomly selected counterparts, which means marriage have no negative or even positive effect in these fields.

Although women's productivity were negatively affected by marriage in short terms, they even publish more in long term of time compared with those randomly selected women scholar. It may because of building a family or having a child drive them to become more productive and organized to achieve both(Ward and Wolf 2004, Joecks, Pull et al. 2013, Lutter and Schröder 2019) and a long trend of fathers becoming more involved in family lives(Sayer, Bianchi et al. 2004).

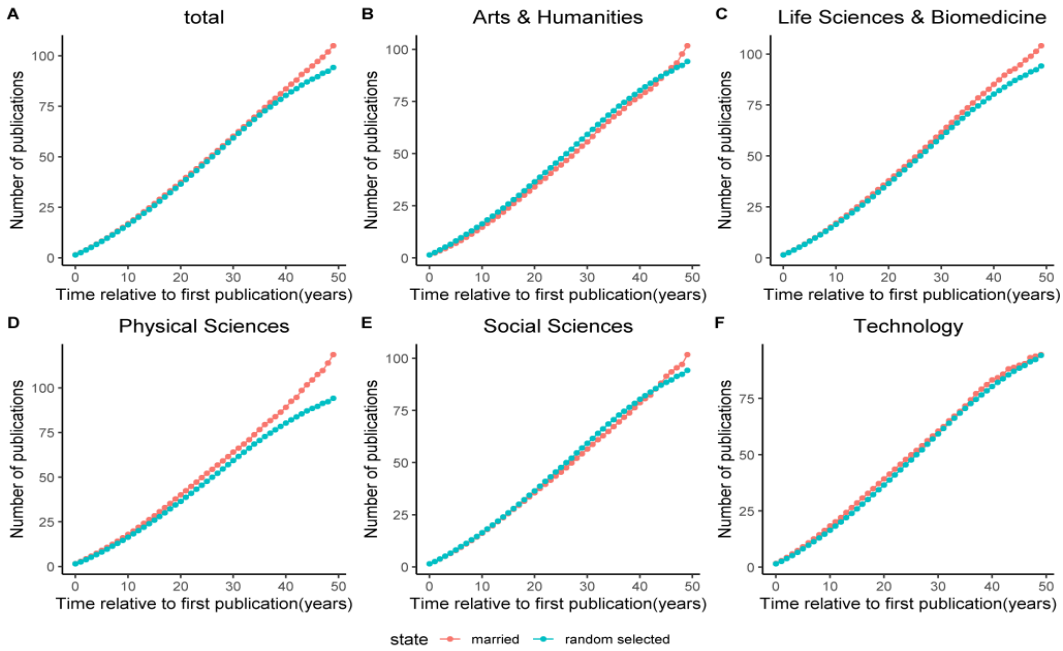


Figure 1. Cumulative number of publications married women and random selected women over their careers. Average cumulative number of publications relative to first publication and productivity gaps for (A) scholar in all field (Nm = 14250, Nr = 12989), (B) scholar in Art & Humanities (Nm = 776, Nr = 521), (C) scholar in Life Sciences & Biomedicine (Nm = 5895, Nr = 6031), (D) scholar in Physical Sciences (Nm = 2141, Nr = 1587), (E) scholar in Social Sciences (Nm = 2673, Nr = 2093), (F) scholar in Technology (Nm = 2082, Nr = 1044).

Further study and limitation

Our next step is to investigate the impact of marriage on career promotion and transitions between academia and industry, using Coarsened Exact Matching (CEM).

We will also incorporate funding data and male scholars as a comparison group to provide a more comprehensive analysis.

A further question that needs to be addressed is: Even if we quantify the impact of marriage on women scholars' productivity and careers, what drives this impact? Do academic women truly prioritize their husbands and families over their academic careers? Or do they still see academia as their primary pursuit but passively experience a decline in productivity due to marriage? Alternatively, do they adopt a "slow accumulation, later breakthrough" career development strategy? What role do their husbands play in this process? To fill these research gaps, we still need more surveys and interviews with women scholars—and perhaps their husbands as well.

The limitations of this study include the following: Since changing one's surname after marriage is a cultural practice specific to certain regions, the regional distribution of married women researchers identified using this method may be uneven. Additionally, with societal changes, even in regions where this practice exists, an increasing number of women choose to retain their original surname after marriage. As a result, the sample of married women researchers obtained may also have temporal limitations. Another limitation is that using the ORCID name change date as the marriage timing for female researchers may not be entirely accurate. Unlike identification documents such as driver's licenses, ORCID is not required for daily life, meaning that the name change recorded in ORCID may lag behind the actual marriage date.

This study enhances our understanding of the relationship between marriage and the productivity and career trajectories of women scholars. By highlighting the complex interplay between marriage, productivity, and career progression, it provides valuable insights for academic institutions and policymakers on how to better support women scholars, ensuring that marriage does not become an insurmountable obstacle to their academic success.

Acknowledgments

The authors declare no conflicts of interest.

References

Cech, E. and M. Blair-Loy (2019). "The changing career trajectories of new parents in STEM." Proceedings of the National Academy of Sciences 116.

Hawks, B. K. and J. Z. Spade (1998). "Women and Men Engineering Students: Anticipation of Family and Work Roles." Journal of Engineering Education.

Joecks, J., K. Pull and U. Backes-Gellner (2013). "Childbearing and (Female) Research Productivity – A Personnel Economics Perspective on the Leaky Pipeline." SSRN Electronic Journal 84(4).

Juraqulova, Z. H., T. C. Byington and J. A. Kmec (2015). "The Impacts of Marriage on Perceived Academic Career Success: Differences by Gender and Discipline." International Journal of Gender, Science, and Technology 7: 369-392.

Lutter, M. and M. Schröder (2019). Is there a motherhood penalty in academia? The gendered effect of children on academic publications, MPIfG Discussion Paper.

Mantai, L. and M. Marrone (2023). "Academic career progression from early career researcher to professor: what can we learn from job ads." Studies in Higher Education 48(6): 797-812.

Mason, M. A. and et al. (2004). "Do Babies Matter (Part II)?" Academe.

Sayer, L. C., S. M. Bianchi and J. P. Robinson (2004). "Are Parents Investing Less in Children? Trends in Mothers' and Fathers' Time with Children." American Journal of Sociology 110(1): 1-43.

Schiebinger, L. and et al. (2010). "Housework Is an Academic Issue." Academe 96(1): 39-44.

Ward, K. and L. Wolf (2004). "Academic Motherhood: Managing Complex Roles in Research Universities." The Review of Higher Education 27: 233-257.

Wolfinger, N. and M. Goulden (2008). "Problems in the Pipeline: Gender, Marriage, and Fertility in the Ivory Tower." The Journal of Higher Education 79.

https://doi.org/10.51408/issi2025_132

Implicit Reporting Standards in Bibliometric Research: What Can Reviewers' Comments Tell Us About Reporting Completeness?

Dimity Stephen[1], Alexander Schniedermann[2], Andrey Lovakov[3], Marion Schmidt[4], Matteo Ottaviani[5], Nikita Sorgatz[6], Roberto Cruz Romero[7], Torger Möller[8], Valeria Aman[9], Stephan Stahlschmidt[10]

[1]*stephen@dzhw.eu*, [2]*schniedermann@dzhw.eu*, [3]*lovakov@dzhw.eu*, [4]*schmidt@dzhw.eu*, [5]*ottaviani@dzhw.eu*, [7]*cruzromero@dzhw.eu*, [8]*moeller@dzhw.eu*, [9]*aman@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Berlin (Germany)

[6]*nikita.sorgatz@hu-berlin.de*
German Centre for Higher Education Research and Science Studies, Berlin (Germany)
Robert K. Merton Center for Science Studies, Humboldt-Universität zu Berlin, Berlin (Germany)

[10]*stahlschmidt@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Berlin (Germany)
Unit of Computational Humanities and Social Sciences (U-CHASS), EC3 Research Group, University of Granada, Granada (Spain)

Abstract

The rapid growth in the number of bibliometric studies in recent years has been accompanied by increasing diversity in the quality of the reporting of these studies' methodologies and results. This ongoing study explores and systematises the quality and completeness of reporting bibliometric research using a bottom-up approach based on open peer review. We first identified 89 bibliometric studies published in library and information science (LIS) journals and conference proceedings and non-LIS journals, and then retrieved the 194 corresponding first-round reviews. From these reviews we extracted 968 reviewer comments pertaining to aspects of reporting the details of these studies, and inductively classified these comments into 11 broad thematic categories and 68 sub-categories. Our preliminary results find that 77% of comments overall and the majority in each broad category were critical, which could be expected given the purpose of peer review to identify opportunities for improvement. In contrast, comments relating to the provision of study data and to the overall assessment of articles were more likely to be positive. The most common themes of reviewers' comments were critically appraising the details of the data, methods, visualisations and tables used, and the clarity of the research questions and text. The finalised results will provide a precise and practical outline of concrete items that should be reported in bibliometric research according to the implicit community standard. Our findings will highlight particular features of bibliometric reporting that could be strengthened, complementing existing initiatives to generate guidance for the complete and accurate reporting of bibliometric studies.

Introduction

Publication output in the field of bibliometrics is growing at an unchecked rate. Larivière (2012) and Jonkers and Derrick (2012) detected a sudden spurt in bibliometric studies in 2003 and growth has only accelerated since then: the number of publications increasing 12-fold from around 800 in 2000-04 to over 10,000 by 2015-19 (González-Alcaide, 2021). Notably, the share of these studies published in library and information science (LIS) journals – the field historically central to

bibliometrics – has steadily decreased over time from around 70% in the 1980s and 1990s to 40% in 2010 (Larivière, 2012) to around 25% in 2019 (González-Alcaide, 2021).

This rapid growth in bibliometric studies may be attributed to several diverse factors. For instance, the prominence of bibliometrics in international, national, and institutional research evaluation and management activities (Cabezas-Clavijo et al., 2023; González-Alcaide, 2021) has raised its profile amongst scholars in all fields. Further, the increasing availability of data sources and analytical software has made bibliometrics accessible to anyone (Cabezas-Clavijo et al., 2023; Boyack, Klavans & Smith, 2022). Viewed cynically, these advances have opened the field to "academic opportunists", who may perceive bibliometric analyses as a quick and easy approach to boosting their publication output (González-Alcaide, 2021). Viewed positively, the self-monitoring capacity in the diverse research fields has been empowered substantially. From either perspective, the prominence and accessibility of bibliometrics has thus generated a wave of interest in our field across disciplines.

While this widespread uptake should be celebrated as an acknowledgment of our field's relevance and potential to contribute broadly to academia, if unchecked, it may also negatively impact the quality, rigour, and development of our field. For instance, our central theories and principles are unlikely to be known to researchers dropping by from other fields to borrow methods and data. Consequently, the bibliometric corpus may be diluted with studies that make minimal contributions to the field or misuse methods and indicators (Jonkers & Derrick, 2012; González-Alcaide, 2021). Individually, such studies are unlikely to have a notable impact on the field. However, in large numbers, they can collectively produce misleading effects, which damages both the theoretical growth of our field and its reputation among academics and policy-makers (Boyack et al., 2022).

Well-documented data and methods are central to the reliability, reproducibility, and robustness of bibliometric studies (Boyack et al., 2022). Evidence of issues in the reporting of bibliometric studies remains currently rather anecdotal. However, a small number of studies that empirically examined reporting quality have found wide variation in the reporting of study characteristics, with good reporting of e.g., search terms, but poor reporting of database characteristics (Koo & Lin, 2023); that substantial numbers of studies lacked the sufficient detail necessary for replicating their findings (Boyack et al., 2022); and that under-reporting of methodological details was widespread in studies both within and outside the LIS field (Cabezas-Clavijo et al., 2023). These findings suggest that the broad community of scholars using bibliometrics could benefit from the guidance in the responsible and effective use of bibliometric data and methods that has long been called for (e.g., Glänzel & Schoepflin, 1994; Glänzel, 1996; González-Alcaide, 2021).

A first step toward providing this guidance is being made with the "Guidance List for repOrting Bibliometric AnaLyses" (GLOBAL) project, which seeks to implement reporting guidelines for bibliometric studies (Ng et al., 2023). GLOBAL comprises a scoping review for existing reporting recommendations and then harnesses the bibliometric community's expertise in developing guideline content.

Establishing and maintaining this continuously evolving shared set of concepts not only facilitates scientific communication, laying the groundwork for progress, but also has the potential to shape education and training in bibliometrics methods. The examination of current reporting standards therefore serves as a critical reflection of our methods, and a consequent broad discussion enables the professional community to agree upon claims for authority and legitimisation and to continue former work to develop the field (American Society for Cell Biology, 2012; Hicks et al., 2015).

Research aims and approach

The aim of this study is to explore and systematise problems in the quality and completeness of reporting bibliometric research. We do so by investigating the question, what reporting issues are identified by peer reviewers in their reviews of bibliometric studies? Our approach is to qualitatively examine peer reviews of bibliometric studies and identify aspects that reviewers raise as well- or poorly reported. For example, reviewers may ask for additional information regarding databases, sample sizes, search terms, filter criteria, or the indicators used, suggesting the provided details were insufficient for understanding or reproducing the study. Instead of pre-defining a set of subjectively ideal reporting criteria, our approach focuses on issues that have been identified by diverse academic peers in open peer review procedures at both central and peripheral bibliometric outlets. As such, our inductive and descriptive approach facilitates a discussion of what features of bibliometrics-based studies the community criticises (or compliments), complementing parallel efforts to jointly define reporting standards in a top-down approach.

Methods

Identification and sampling of bibliometric studies

The methods applied in the study are shown in Figure 1. We used a two-step process to identify bibliometric studies in journals and conference proceedings. We defined a bibliometric study as a study that used a bibliometric data source (e.g., WoS, Scopus) and one or more metadata fields (e.g., journal, discipline) to compare two or more entities or groups (e.g., authors, institutions, countries) to contribute knowledge to its field (e.g., one database covers more journals than another or one institution is more productive than another).

Our sampled studies included articles in LIS and non-LIS journals, and submissions to a LIS conference. As noted, a substantial number of bibliometric studies now appear in non-LIS journals and these may be reviewed by peers less familiar with the details necessary to sufficiently report a bibliometric study than reviewers of articles in LIS journals, which potentially increases the diversity of aspects raised. Similarly, conference papers are usually shorter in length than articles and may contain fewer methodological details and results than articles, and so reviewers may highlight particularly important features when these are missing. This sample of both articles and conference submissions may thus capture a wide array of issues raised by reviewers, aligning with the interdisciplinarity present in bibliometrics.

We first required a sample of bibliometric studies in each category (NLIS article, LIS article, LIS conference submission) with open peer reviews. The 27th International Conference on Science, Technology and Innovation Indicators (STI 2023) used an open peer review process, so we used these submissions to represent conference proceedings (sample A). To identify bibliometric studies in these submissions, we extracted the title, abstract and keywords (TAK) of each submission from the Orvium website using the rvest (Wickham, 2022a), tidyverse (Wickham, et al. 2019), jsonlite (Ooms, 2014), and stringr (Wickham, 2022b) R packages. We then narrowed the submissions to those that used a bibliometric data source by searching for any of the following (case insensitive) terms in the TAK: Web of Science, WoS, Scopus, Dimensions, Openalex, Open Alex, Pubmed, Crossref, SciELO, Wikidata, Overton, altmetric, bibliometric data, DOAJ. We then manually screened the full-texts of these submissions to assess whether they fulfilled the aforementioned criteria of using a metadata field and comparing groups to make a knowledge claim, and retained those that fulfilled these criteria as bibliometric studies.

To identify N/LIS articles with open peer reviews, we performed two searches of the online WoS database on 29 February 2024, including the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. First, to identify LIS articles (sample B), we searched for any of the aforementioned bibliometrics data sources in the Topic (TAK) field. In addition, we restricted the publication years to 2018-2024, the WoS Subject Category to "Information & Library Science", the document type to article, and filtered the results to those articles that were open access (OA) and had open peer reviews available. We performed the same search to identify non-LIS articles (sample C), with the following changes: Category was not "Information & Library Science", the title did not include “Protocol”, Dimensions and Pubmed were removed, and “scientometric” was added as a search term, as we observed authors to use bibliometric and scientometric interchangeably. “Protocol” was excluded to remove study protocols. Dimensions was removed because it is unlikely to refer to the database outside of LIS, and Pubmed was removed as its inclusion returned many out-of-scope systematic reviews. For both samples, we then manually screened the studies’ full-texts to retain those that fulfilled our criteria as bibliometric studies. We then downloaded the first-round peer reviews for all bibliometric studies via the “Open Peer Reviews” link in WoS.

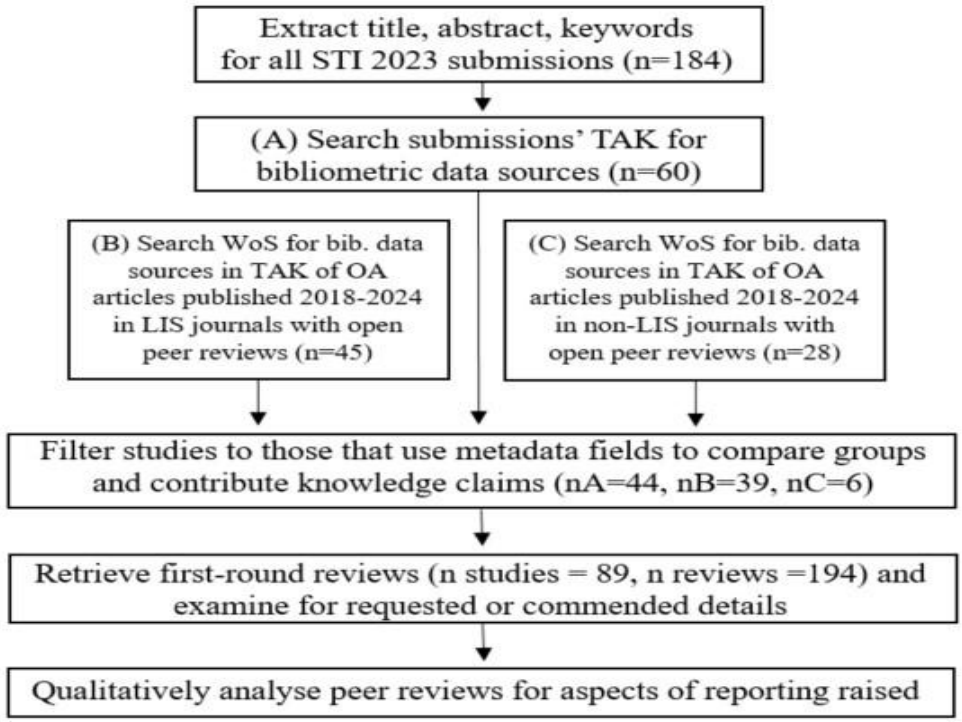


Figure 1. Flowchart of method.

Qualitative analysis of reviewer comments

To prepare for our qualitative analysis of the peer reviewers' comments, we first extracted all comments pertaining to reporting a bibliometric study from the reviewers' reports for all three samples. In this process, each team member examined and extracted comments from approximately 20 peer reviews. The comments could be positive, such as praise for the clear or detailed description of the methodology; negative, such as critiquing the study's limitations; or neutral, such as suggestions for additional references. At this stage, we aimed to collect as much information as possible and filter out irrelevant information later in the analysis.

We then categorised the comments into broad themes based on the overarching concept of the comment. Here, in a group process, we discussed the comments' focus and identified and allocated comments to one or more high-level categories. To enhance the specificity of the concepts addressed, we as a group then further assessed the comments in each category and identified a set of more specific sub-categories. For instance, the reviewer comment "What is the unit for y-axis in Figure 7?" was first assigned to the broad category of *Visualisations and Tables* and then sub-categorised to *(Un)clear presentation*. Sub-categories were named neutrally as comments could be positive, neutral, or negative. Once classified, all comments in each sub-category were reviewed for consistency and reclassified to other or new sub-categories as required. In this way, we inductively classified all comments to both broad categories and more specific sub-categories based on the concept addressed in the comment.

Results

The total sample examined consisted of 194 reviews of 89 bibliometric studies: 11 reviews of 6 studies published in NLIS journals, 79 reviews of 39 studies published in LIS journals, and 104 reviews of 44 LIS conference papers. The LIS articles were all published in Quantitative Science Studies, as the only WoS-indexed LIS journal with open peer review. The NLIS studies were published in six journals: Ecological Solutions and Evidence, Engineering Reports, Environmental Research Letters, Internet Technology Letters, Journal of Oral Rehabilitation, and Royal Society Open Science. The low number of NLIS studies occurs as the open peer review restriction severely limited the sample. On average, reviews of conference papers were 287 words in length (range = 31-1,091 words), which was – as could be expected – shorter than article reviews. Reviews of bibliometric studies in the LIS journals were notably longer (mean = 710 words, range = 76-2,605) than articles in NLIS journals (mean = 536 words, range = 35-2,062).

The initial coding of the reviews identified 1,030 relevant comments. Sixty-two comments were later deemed to be out of scope of the analysis and removed, leaving 968 comments in scope. The first classification process identified 11 broad themes: *Clarity and validity of concepts*; *Clarity of presentation*; *Description of data/methods*; *Description of results*; *Visualisations and tables*; *Limitations*; *Conclusions*; *Open Science/Reproducibility*; *Declarations*; *Links to literature/references*; and *Overall assessment*. The second classification process identified 68 sub-categories of these themes. Table 1 shows the number and percentage of comments in the 11 broad categories and the number and percentage of each category's comments that were negative (i.e. critical of the manuscript), neutral, or positive. As comments could be classified to more than one category, the total count of comments exceeds 968.

Table 1. The number and percentage of comments in the 11 broad categories and the number and percentage of comments in each category that were positive, neutral, or negative, ordered by the total number of comments.

Category	*No. (%) comments*	*No. (%) negative*	*No. (%) neutral*	*No. (%) positive*
Description of data / methods	329 (29.1)	287 (87.2)	5 (1.5)	37 (11.2)
Clarity of presentation	139 (12.3)	89 (64.3)	0 (0.0)	50 (35.7)
Visualisations and tables	136 (12.0)	118 (86.8)	6 (4.4)	12 (8.8)
Description of results	131 (11.6)	111 (84.7)	11 (8.4)	9 (6.9)
Overall assessment	118 (10.5)	51 (42.9)	0 (0.0)	67 (56.8)
Links to literature / references	83 (7.4)	69 (84.1)	2 (2.4)	12 (14.5)
Clarity and validity of concepts	62 (5.5)	57 (91.9)	0 (0.0)	5 (8.1)
Conclusions	59 (5.2)	54 (91.5)	0 (0.0)	5 (8.5)
Open science / reproducibility	42 (3.7)	20 (47.6)	0 (0.0)	22 (52.4)
Limitations	29 (2.6)	20 (69.0)	2 (6.9)	7 (24.1)
Declarations	1 (0,1)	1 (100.0)	0 (0.0)	0 (0.0)
Total	1,129 (100)	877 (77.7)	26 (2.3)	226 (20.0)

Nearly a third of reviewers' comments pertained to the authors' description of the data and or methods used in the study (329, 29.1%), the majority of which (87.2%)

were critical, while 11.2% of comments praised the methodological information presented. The next most common comments regarded the clarity of the information presented (12.3%), and the visualisations and tables used (12.0%). In the former category around two-thirds of comments were critical of, for instance, the clarity of research questions and the structure of the text, while one-third of comments regarded these features positively. Comments regarding the content and presentation of visualisations and tables, however, were largely critical (86.8%). Similarly, overall, 77.7% of comments were critical of the manuscripts' reporting, which aligns with the aim of peer review to identify potential issues and suggest improvements to the authors. In contrast, comments relating to open science/reproducibility (e.g., the provision of the data and or scripts used in the study) and the overall assessment of the study (e.g., its contextualisation in the existing literature, appropriateness of its design to address the research question, and its originality, utility, and relevance) were more often positive than negative. However, this latter instance may have been influenced by the fact that all articles examined were eventually accepted for publication.

These preliminary results provide initial insights into the issues raised and details praised by reviewers of bibliometric studies. This study is ongoing and we intend to finalise the qualitative analysis of the reviewers' comments, particularly the sub-category level, which will provide greater granularity of the themes discussed in the comments and highlight specific aspects of the reporting of bibliometric studies that should be addressed by authors. Further, we plan to compare the theme and prevalence of comments between articles and conference submissions and between NLIS and LIS articles to investigate potential differences in reviewers' focus or authors' reporting between groups. Finally, we plan to distill the results into a precise and practical list of concrete items that should be reported in bibliometric research according to the implicit community standard, and present this for discussion at the conference.

We anticipate that our results will provide a descriptive and inductive perspective of the aspects of reporting bibliometric studies raised by peer reviewers. This will highlight particular features of bibliometric reporting that could be strengthened and complement initiatives such as GLOBAL, which take an expert-based top-down approach to generating guidance in complete and accurate reporting of bibliometric studies. The availability and up-take of such guidance could enhance the reliability, reproducibility, and robustness of bibliometric studies.

References

American Society for Cell Biology. (2012). San Francisco declaration on research assessment (DORA). https://sfdora.org/

Boyack, K. W., Klavans, R., & Smith, C. (2022). Raising the bar for bibliometric analysis. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), 26th International Conference on Science and Technology Indicators, STI 2022. sti22143. DOI: 10.5281/zenodo.6975632.

Cabezas-Clavijo, A., Milanés-Guisado, Y., Alba-Ruiz, R., & Delgado-Vázquez, A. M. (2023). The need to develop tailored tools for improving the quality of thematic bibliometric analyses: Evidence from papers published in Sustainability and

Scientometrics. Journal of Data and Information Science, 8(4), 10–35. DOI: 10.2478/jdis-2023-0021.
Glänzel, W. (1996). The need for standards in bibliometric research and technology. Scientometrics, 35(2), 167–176. DOI: 10.1007/BF02018475.
Glänzel, W., & Schoepflin, U. (1994). Little scientometrics, big scientometrics... and beyond? Scientometrics, 30(2–3), 375–384. DOI: 10.1007/BF02018107.
González-Alcaide, G. (2021). Bibliometric studies outside the information science and library science field: uncontainable or uncontrollable? Scientometrics, 126(8), 6837–6870. DOI: 10.1007/s11192-021-04061-3.
Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. & Raflos, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520, 429–431. https://www.nature.com/articles/520429a.
Jonkers, K. & Derrick, G. (2012). The bibliometric bandwagon: Characteristics of bibliometric articles outside the field literature. Journal of the American Society for Information Science and Technology, 63(4), 829-836. DOI: 10.1002/asi.22620.
Koo, M., & Lin, S-C. (2023). An analysis of reporting practices in the top 100 cited health and medicine-related bibliometric studies from 2019 to 2021 based on a proposed guidelines. Heliyon, 9(6), e16780. DOI: 10.1016/j.heliyon.2023.e16780.
Larivière, V. (2012). The decade of metrics? Examining the evolution of metrics within and outside LIS. Bulletin of the American Society for Information Science and Technology, 38(6), 12-17. DOI: 10.1002/bult.2012.1720380605.
Ng, J. Y., Haustein, S., Ebrahimzadeh, S., Chen, C., Sabe, M., Solmi, M., & Moher, D. (2023). Guidance List for reporting bibliometric analyses (GLOBAL): A research protocol. Open Science Framework. DOI: 10.17605/OSF.IO/MTXBF.
Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between JSON data and R objects. arXiv. DOI: 10.48550/arXiv.1403.2805.
Wickham, H. (2022a). rvest: Easily Harvest (Scrape) Web Pages. (R package version 1.0.3). https://CRAN.R-project.org/package=rvest.
Wickham, H. (2022b). stringr: Simple, Consistent Wrappers for Common String Operations. (R package version 1.4.1). https://CRAN.R-project.org/package=stringr.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R, et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

https://doi.org/10.51408/issi2025_131

Is Journal Citation Indicator A Good Metric for Art & Humanities Journals Currently?

Yu Liao[1], Li Li[2], Zhesi Shen[3]

[1] liaoyu@mail.las.ac.cn, [2] lili2020@mail.las.ac.cn, [3] shenzhs@mail.las.ac.cn
National Science Library, Chinese Academy of Science, Beijing (China)

Abstract

Probably Not.

Journal Citation Indicator (JCI) was introduced to address the limitations of traditional metrics like the Journal Impact Factor (JIF), particularly its inability to normalize citation impact across different disciplines. This study reveals that JCI faces significant challenges in field normalization for Art & Humanities journals, as evidenced by much lower correlations with a more granular, paper-level metric, CNCI-CT. A detailed analysis of Architecture journals highlights how journal-level misclassification and the interdisciplinary nature of content exacerbate these issues, leading to less reliable evaluations. We recommend improving journal classification systems or adopting paper-level normalization methods, potentially supported by advanced AI techniques, to enhance the accuracy and effectiveness of JCI for Art & Humanities disciplines.

Introduction

The Journal Impact Factor (JIF) has long been the predominant metric for evaluating journals, celebrated for its simplicity and widespread acceptance (Miles et al., 2018). However, its limitations have been widely criticized, including its failure to account for variations in citation potential across disciplines (Althouse et al., 2009; Nederhof 2006), differences in document types, the constraints of a short citation window, and the impact of highly skewed citation distributions (Larivière and Sugimoto, 2019; Bordonset al., 2002). To address some of these issues, the Journal Citation Indicator (JCI) was introduced. JCI calculates the average Category Normalized Citation Impact (CNCI) of articles published in a journal, normalized using a journal-level subject category classification system (hereafter referred to as JCI-WoS).

In recent years, JCI has gained traction as a metric, especially for evaluating Art & Humanities journals, which face unique challenges due to their distinctive citation practices and field-specific characteristics (Torres-Salinas et al., 2022). Beginning with the 2023 Journal Citation Reports (JCR), Clarivate Analytics adopted JCI-based quartiles, replacing JIF-based quartiles, for Art & Humanities journals. Although prior studies have demonstrated a strong correlation between JCI and JIF for journals indexed in SCIE and SSCI, the performance of JCI as a field-normalization metric for Art & Humanities journals remains insufficiently examined.

This study evaluates JCI's effectiveness for Art & Humanities journals by comparing it with CNCI calculated based on Citation Topics (hereafter referred to as CNCI-

CT), a more granular, paper-level classification system. A high correlation between JCI and CNCI-CT indicates that JCI has effectively achieved field normalization. Specifically, this study addresses the following research questions:

1. Does the correlation between JCI and CNCI-CT for Art & Humanities journals differ from that observed for Science and Social Science journals?
2. If differences exist, what underlying factors contribute to these discrepancies?

Through this investigation, we aim to provide a detailed evaluation of JCI's field-normalization performance in Art & Humanities journals, offering valuable insights into its appropriateness as a standard metric for these disciplines.

Data and Methods

To evaluate the performance of the JCI in the context of Art & Humanities journals, we obtained the CNCI values of 22,979 journals indexed in SCIE, SSCI, AHCI, and ESCI from the InCites database in December 2024. Only documents categorized as articles and reviews published during the 2021–2023 period are included. For each journal, two CNCI values were extracted:

1. **CNCI based on Subject Category (JCI):** This is calculated at the journal level by normalizing the citation impact of articles against all other documents within the same journal's subject category as defined by the Web of Science. JCI is essentially the average CNCI-WoS for a journal.
2. **CNCI based on Citation Topics-meso level (CNCI-CT):** This is calculated at the paper level by normalizing citation impact based on a more granular, hierarchical classification system called Citation Topics. The meso level topics is selected as it has similar granurity with subject category

Results

RQ1: Does the correlation between JCI and CNCI-CT for Art & Humanities journals differ from that observed for Science and Social Science journals?

A1: Low correlations between JCI and CNCI-meso are found for Art & Humanity related subject categories, implying JCI's in-effectiveness in citation field normalization.

Figure 1 illustrates the correlation coefficients between JCI and CNCI-CT across various subject categories. A higher correlation indicates a closer alignment between the two metrics for journals within a given category. We grouped subject categories into three broad groups: Science, Social Science, and Art & Humanities.

As shown in Figure 1, the Science and Social Science groups exhibit consistently high and tightly clustered correlation coefficients, reflecting strong alignment between JCI and CNCI-CT. In contrast, the Art & Humanities group displays a wider

range of lower correlation coefficients. Notably, categories such as *Architecture* and Theater show the weakest correlations. This discrepancy indicates that there is a significant divergence between journal-level classification-based and paper-level classification-based normalization metrics for journals in the Arts & Humanities, raising concerns about the effectiveness of JCI in normalizing citation disparities within these subject categories.

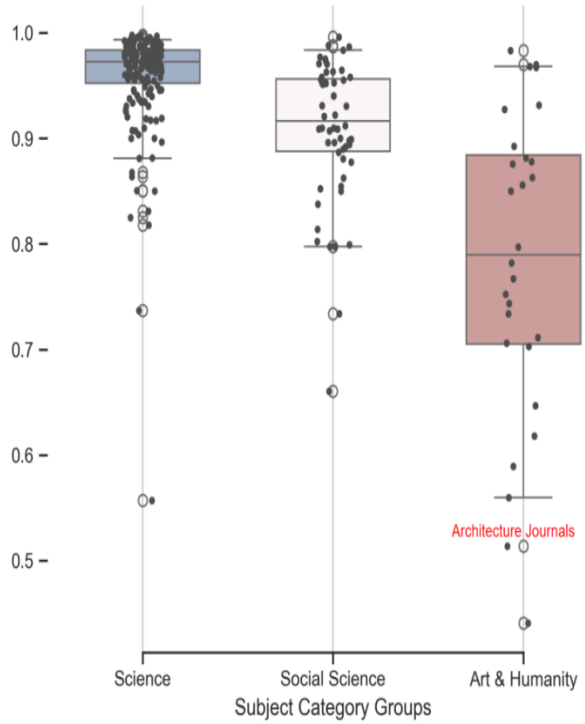


Figure 1. Correlation coefficients between JCI and CNCI-CT for subject categories. Each dot represents a subject category, with its value indicating the correlation coefficient. Categories are grouped into three clusters: Science, Social Science, and Art & Humanities.

RQ2: If differences exist, what underlying factors contribute to these discrepancies?
A2: Through the case study of architecture journals, it was revealed that some art & humanities journals, despite publishing a significant number of science or social science papers, are not co-assigned to the science and social science categories. This omission results in these journals gaining a substantial advantage in the JCI.
Among Art & Humanities categories, *Architecture* demonstrates one of the lowest correlations between JCI and CNCI-CT. To further investigate, we analyzed the structure of this category in detail.

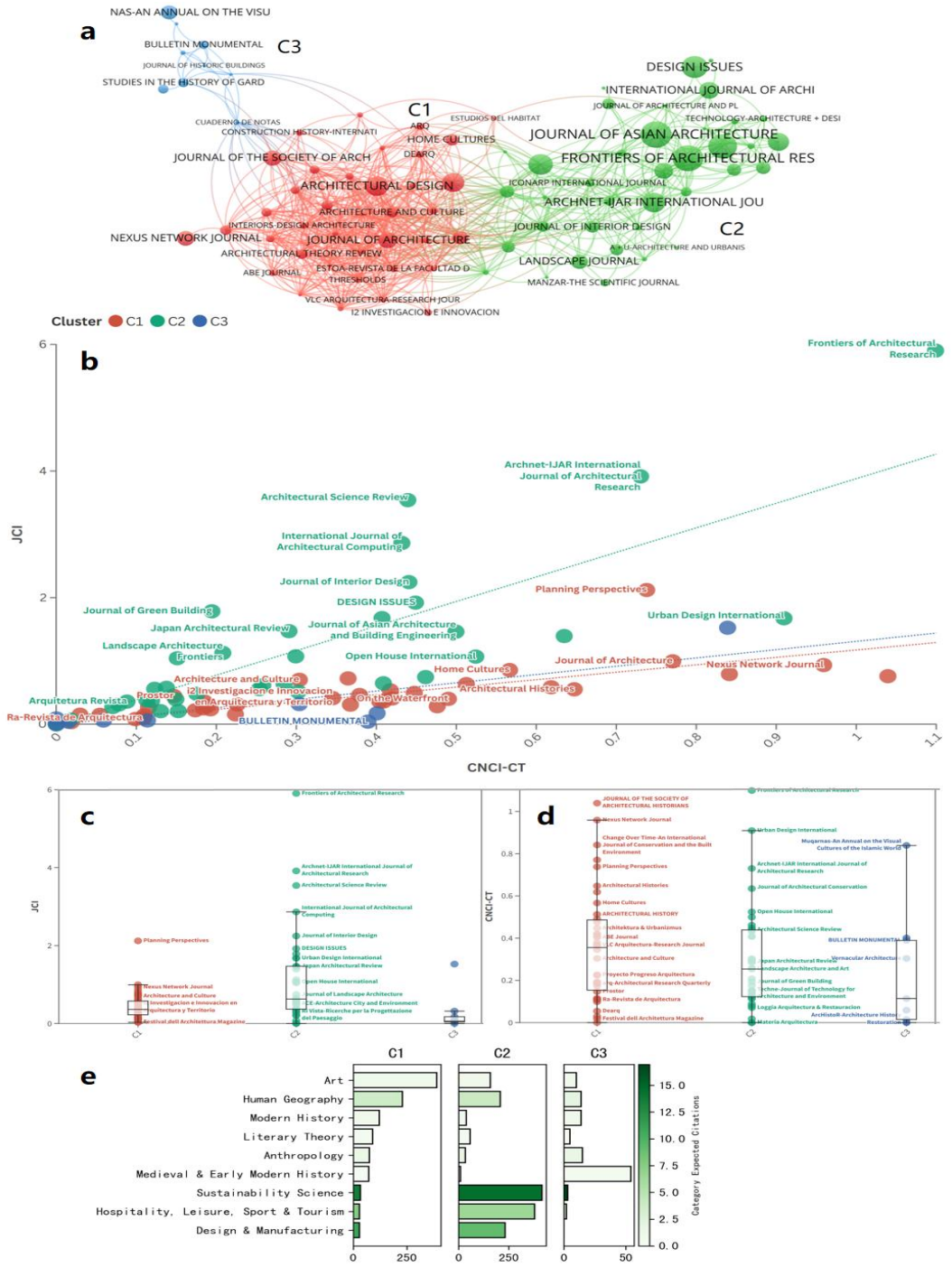


Figure 2. Field normalization performance for Architecture journals.

(a) Similarity network of Architecture journals, showing three identified clusters. (b) Scatter plot comparing JCI and CNCI-CT values for journals in each cluster. (c) Distribution of JCI values by cluster. (d) Distribution of CNCI-CT values by cluster. (e) Distribution of covered citation topics of each cluster with color representing the category expected citation.

Cluster Analysis

Figure 2(a) presents the similarity network of Architecture journals, where journals publishing similar content are positioned closer together. Using a community detection algorithm, we identified three distinct clusters.

Disparities in JCI and CNCI-CT Across Clusters

Figures 2(c) and 2(d) compare the distributions of JCI and CNCI-CT values across the three clusters. As shown in Figure 2(c), Cluster 2 (green dots) exhibits significantly higher JCI values compared to Clusters 1 and 3. However, Figure 2(d) reveals that CNCI-CT values are more evenly distributed across all three clusters. The scatter plot in Figure 2(b) highlights the substantial advantage that JCI provides to journals in Cluster 2, suggesting that JCI does not fully account for citation disparities within the *Architecture* category.

Content Differences Across Clusters

An examination of publication topics in Cluster 2 journals reveals a higher proportion of articles related to *sustainability science* topics with higher citation potential compared to traditional Architecture topics, as shown in Fig.2(e). Despite this interdisciplinary content, most Cluster 2 journals remain solely classified under the *Architecture* category, with only a small number being co-classified into science or social science categories.

Implications for Field Normalization

Because JCI uses journal-level subject category normalization, Cluster 2 journals benefit significantly from their inclusion in a single, less-cited category, despite publishing content that overlaps with higher-citation Science fields. In contrast, CNCI-CT employs paper-level normalization based on Citation Topics, which more effectively captures thematic and disciplinary diversity, resulting in a more balanced evaluation of journals across clusters.

Conclusions and Discussions

This study examines whether the Journal Citation Indicator (JCI) effectively addresses field normalization challenges for Art & Humanities journals. By comparing JCI with CNCI-CT, a field-normalized indicator based on paper-level classification, we find significantly lower correlations between the two metrics in Art & Humanities categories. This indicates that JCI currently struggles to handle field normalization disparities in these fields.

A detailed analysis of *Architecture* journals reveals that this issue primarily arises from journal-level misclassification. Similar patterns are observed in other Art & Humanities categories, such as *Art* and *Religion*. Due to the lower citation density

characteristic of Art & Humanities compared to Science and Social Science fields, the effects of misclassification are more pronounced, further reducing the reliability of JCI in these areas.

To address these limitations, we recommend prioritizing the optimization of journal classifications (Yu et al., 2025) before expanding the use of JCI. Alternatively, adopting a paper-level classification system for field normalization (Sichao et al., 2023) could provide a more robust solution. However, implementing paper-level classification in Art & Humanities faces unique challenges: approximately 20% of papers in these fields are not assigned citation topics, compared to nearly 0% in Science and Social Science. To overcome these challenges, advanced AI methods, including large language models (LLMs), could be employed to assign citation topics based on titles and abstracts. These tools have the potential to improve classification coverage significantly, enhancing the accuracy of field normalization and making metrics like JCI more reliable for evaluating Art & Humanities journals.

References

Althouse, B.M., West, J.D., Bergstrom, C.T., Bergstrom, T.(2010). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science & Technology*, 60 (1), 27–34.

Bordons, M., Fernández, M., & Gómez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, 53(2), 195–206.

Larivière, V., Sugimoto, C.R. (2019). The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects. In: Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M. (eds) Springer Handbook of Science and Technology Indicators. Springer Handbooks. Springer, Cham.

Miles, R. A., Konkiel, S. & Sutton, S., (2018). Scholarly Communication Librarians' Relationship with Research Impact Indicators: An Analysis of a National Survey of Academic Librarians in the United States. Journal of Librarianship and Scholarly Communication, 6(1), eP2212.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. Scientometrics, 66(1), 81–100.

Sichao Tong, Fu-You Chen, Liying Yang, Zhesi Shen, Novel utilization of a paper-level classification system for the evalution of journal impact: an update of the CAS Journal Ranking. Quantitative Science Studies (2023) 4(4):960-975

Torres-Salinas, D., Valderrama-Baca, P., & Arroyo-Machado, W. (2022). Is there a need for a new journal metric? Correlations between JCR Impact Factor metrics and the Journal Citation Indicator—JCI. Journal of Informetrics, 16(3), 101315. Redirecting

Yu Liao, Jiandong Zhang, Liying Yang, Zhesi Shen. (2025). The initiative of refining CAS journal subject classfiication system. Journal of Data and Information Science, 1-3.

https://doi.org/10.51408/issi2025_124

Mapping and Quantifying the Boundaries in Research Data Sharing based on Data Policy

Yizhan LI [1], Mingze ZHANG [2], Lu DONG [3], Zexia LI [4]

[1] liyz@mail.las.ac.cn, [2] zhangmingze@mail.las.ac.cn, [3] donglu@mail.las.ac.cn, [4] lizexia@mail.las.ac.cn

National Science Library, Chinese Academy of Sciences, Beijing (China)

Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

Abstract

Balancing research data openness with security concerns necessitates regulatory constraints, yet the absence of standardized quantitative thresholds complicates cross-institutional and cross-border data sharing. This study examines 72 policy documents from the US, EU, and UK. Using a large language model (LLM)-based prompt engineering approach, we extract and quantify data-sharing constraints through a two-stage framework: (1) Constraint Identification, detecting access limitations, and (2) Quantitative Relation Extraction, identifying key metrics such as data scale, durations, etc. Our findings categorize data-sharing boundaries into three types: mandatory restrictions (red line), conditional constraints (blue line), and ambiguous areas shaped by evolving technologies. A comparative analysis of key quantitative constraints, like embargo periods, reveals inconsistencies across policies, highlighting the need for regulatory alignment. Additionally, we identify subject-specific access restrictions that resemble controlled data list. Future research will refine constraint mapping, analyze policy evolution, and explore interdisciplinary data governance. These efforts aim to enhance policy clarity, enhance operational efficiency, and support international research collaboration.

Introduction

The rapid growth and large-scale accumulation of research data have shifted it from being a mere byproduct of research activities to a foundational resource for scientific investigation. Fields such as earth sciences, life sciences, materials science, and computer science increasingly exemplify the defining features of data-intensive knowledge discovery. This transformation has been propelled by initiatives like *the Global Open Science Movement* and *the Fourth Paradigm of Scientific Research*, which emphasize open access and data-driven discoveries. These efforts have enabled the unprecedented reuse and interconnection of geospatial, ecological, personal sensitive, health, and agricultural data (Xiang & Cai, 2021; George, 2019). However, this openness introduces significant challenges, including risks to security, personal privacy, intellectual property rights, commercial interests, and ethics (Li et al., 2023; Amiri-Zarandi et al., 2022; Majeed, 2021; Zigomitros et al., 2020). These issues, exacerbated by the rapid development of emerging and disruptive technologies, underscore the growing importance of research data security. Nations also have faced fundamental disagreements over principles governing the cross-border flow of data, further complicating efforts to safeguard data (Ducato, 2020).

To address these challenges, national laws establish overarching guidelines, while major funding agencies, research institutions, and international scientific programs implement policies to regulate the sharing and use of research data. These measures

aim to mitigate security risks by creating a multi-tiered framework of regulations and intangible boundaries that define the flow and usage of data. The unique characteristics of research data, such as shareability, non-exclusivity, asymmetry, transferability, long-term accumulation, and its public interest nature—further complicate the balance between openness and security (Li et al., 2024). These characteristics result in diverse priorities and roles for national authorities, funding agencies, researchers, and data contributors within the data sharing and value chain(Li et al., 2022).

In real-world contexts, constraints on research data sharing are often principle-based, with sensitive data classified primarily by the harm or loss they may cause. For researchers, such guidelines often lack practical applicability. While some rules employ quantitative metrics and thresholds, these face challenges such as inconsistent standards and thresholds that evolve with technological advancements and shifting risk factors. This paper focuses on research data sharing policies and seeks to address the following questions:

Q1. What are the current boundaries of research data sharing, and in what forms or manifestations do they appear?

Q2. Can the boundaries of research data sharing be quantitatively defined?

By combining policy text analysis with a quantitative framework, this paper aims to bridge the gap between principle-based and operational rules. This approach enables researchers to navigate data-sharing complexities with greater clarity, consistency, and security. Additionally, it standardizes guidance and fosters a benchmark for dialogue across institutions, organizations, and countries.

Dataset Construction, Processing and Methodology

Research data policy collection and its metadata

This study provides a systematic examination of the legal frameworks and regulatory instruments governing research data across three jurisdictions: the United States (US), the European Union (EU), and the United Kingdom (UK). A total of 72 policy documents were comprehensively collected from publicly accessible sources (Table SI-1), with corresponding metadata and access information recorded (Table 1). These documents encompass a broad range of national legislative acts, directives, regulations, rules, guidance materials, and executive orders related to the data domain in the US, EU, and UK. While not all documents specifically target research data, it is evident that research data—as a critical subset of broader data ecosystems—must adhere to these overarching policies, particularly with respect to data sharing and security. The corpus also includes strategic policy documents that outline anticipated developments and policy trajectories for data sharing in the coming years. In addition, the study reviews data management requirements issued by major funding agencies (e.g., the US National Science Foundation and UK Research and Innovation), prominent research institutions, and international scientific collaboration initiatives. These requirements frequently reflect disciplinary particularities and address diverse data modalities, including text, tables, images, and audio.

Table 1. Metadata of policy documents related with research data sharing topic.

Field Name	*Description*
OID	Unique ID
File Name	The official name of the document
Type	The type of policy document, including *Act*, *Directive*, *Regulation*, *Rule*, *Strategy*, *Guidance*, etc.
SubType	*Rule* is subdivided into *Rules_Government*, *Rules_Sponsor_Public*, *Rules_Sponsor_Private*, *Rules_Project*, *Rules_Institution*, *Rules_International Organization*, *Rules_International Project*, etc.
Issuing Authority	The name of the organization that issued the document
Country/Region	The geographical scope where the document applies
Issuing Date	The official issuing date of the document
Enforceability	Mandatory or not
Access Address	URL or PDF file download from the official website
Policy Language	English, etc.

Paragraph extraction and analysis with LLM

In the field of policy informatics, several foundational studies have outlined common methods and procedures for the quantitative analysis of policy texts. Automated policy text analysis typically involves three main tasks—classification, clustering, and scaling(Grimmer & Stewart, 2013)—and follows a general workflow that includes preprocessing, stemming, bag-of-words model, category development and coding, reliability and validity checks, and content interpretation(Cao & Zhang, 2022; Bardach & Patashnik, 2019; Lucas et al., 2015). These methods have been applied to various types of policy documents, such as legislative acts and international treaties(Yang et al., 2020), often focusing on entities or clauses as the unit of analysis. In recent years, the emergence and widespread adoption of large language models have made policy text analysis more streamlined and fine-grained. This study employs a structured prompt engineering methodology, integrating template construction and iterative optimization to extract policy constraints on research data sharing (Figure 1). Using LLMs like ChatGPT-4o and DeepSeek-R1, we propose a two-stage framework: (1) Constraint Identification – domain-adapted prompts guide LLMs to detect data-sharing restrictions (e.g., access limits, usage boundaries); (2) Quantitative Relation Extraction – refined templates identify constraint-related metrics (e.g., temporal restrictions, user quotas). Our prompt engineering follows the "Role-Objective-Skill-Workflow-Constraint-Output" framework (Figure SI-1). A test set (20% of 72 policy documents) was iteratively optimized, with representative policies selected from different jurisdictions, policy types (e.g., Act, Directive) (Caufield et al., 2024; Chen et al., 2024; Durmaz et al., 2024; Yang et al., 2024).

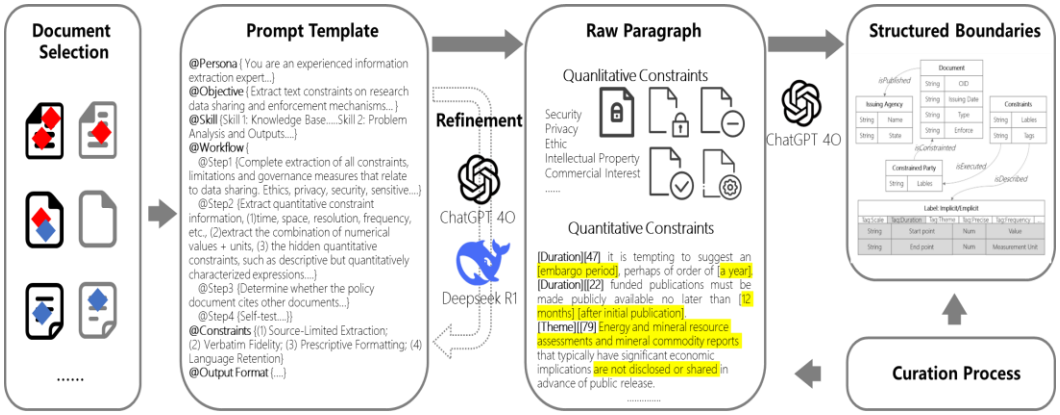


Figure 1. Overview of paragraph extraction process.

All policy documents used in this study are publicly available and contain no personal or sensitive information. Nonetheless, the use of LLMs for data extraction raises concerns about output accuracy and interpretability. To address these issues, a curation process—implemented as a human-in-the-loop review—was used to manually verify and refine all extracted results, ensuring their reliability.

Discussion

Conceptual description and Forms of boundaries

The analysis of policy texts reveals that the openness and sharing of research data are subject to certain boundaries. These boundaries vary depending on the type of research data governance documents and the nature of the institutions that issue them. They can be categorized into three types: mandatory boundaries (Red line), conditionally negotiable boundaries for targeted sharing (Blue line), and areas of uncertainty that remain undefined (Figure 2).

Red line: This category includes confidential data related to national security, data sovereignty, and personal privacy, which are clearly defined by national or regional laws, regulations, and confidentiality agreements.

Blue line: This category refers to data that can be shared under specific conditions, such as restrictions on the use of research data in particular network environments, among defined user groups, or within a controlled scope of access.

Ambiguity area: This category pertains to areas that are still under debate or evolving alongside technological advancements. For example, the development of gait recognition technology allows surveillance data from public spaces to be used for identifying individuals based on gait features. As a result, this data has been classified as personal information and recognized as a form of biometric data,

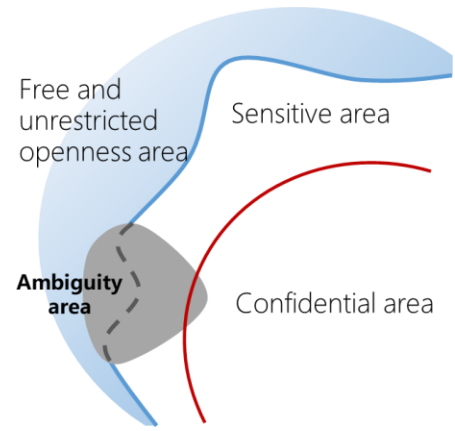


Figure 2. Three forms of research data sharing boundaries.

similar to fingerprints or voiceprints. This is a typical case of how advancements in technology lead to changes in data sensitivity, resulting in a contraction of data sharing boundaries.

Spectrum of boundaries

Figure 3 presents the relationship between the classification of 72 policy documents and the defined boundary constraints, along with the document types and key elements of these constraints. The red-to-blue gradient denotes mandatory regulatory changes, whereas the green-to-yellow gradient represents a shift from qualitative to quantitative constraints.

Mandatory legal regulations typically prioritize qualitative, principle-based constraints. For instance, research data sharing is generally governed by principles such as national security, ethics, privacy protection, and intellectual property rights, etc. Moreover, certain parameters may be subject to principle-based restrictions, meaning that while requirements such as assessments and reviews for large-scale data sharing are imposed, specific quantitative thresholds are not explicitly defined. However, current document analyses indicate that explicit quantitative thresholds are seldom specified, underscoring the need for supplementary regulatory frameworks.

Each NO. represents a document, and the colors of the NO.s represent different countries /region. Legend US, EU, UK, Multi		Red Line			Blue Line					Ambiguity Area	
					Rule						
		Act	Regulation	Directive	Governament rules	Sponsor rules	Project rules	Institution rules	International Organization rules	Guidance	Strategy
1. Qualitative Constraints		58, 61,	33,	62,	45, 65, 72,	01, 13, 14, 15, 17, 44, 52, 53, 56, 57,			26, 29,	42,	27, 43,
2.Quantitative Constraints	2.1 Implicit	24, 25, 69, 70,	23,	32,	08, 68	12, 18, 71,	19, 20,		39,		28,
	2.2 Explicit		30, 31,	51,	04, 05, 09, 21, 22, 40,	02, 03, 06, 10, 11, 16, 38, 46, 47, 48, 50, 59, 60, 63, 66, 67,	64,	36,	34, 35, 54, 55, 65,	41,	
	Scale	24, 25, 69, 70,	23,			38, 66,					
	Precise	24, 25,	23,		08,	66,			55, 65,		
	Frequency	24, 25, 69, 70,		32,	68,	60,			65,		
	Duration	69, 70,		51,	04, 05, 09, 21, 22,	02, 03, 06, 10, 11, 12, 16, 18, 38, 46, 47, 48, 59, 63, 67	19, 20	36,	39, 54, 65		28,
	Scope					46, 47, 60,			35, 55, 65		
	Data Source					46, 47, 66,	64,		65,		
	Theme	24, 25, 69, 70,	23,			16, 49, 50, 59, 63,	64,		34, 54, 55, 65	41,	
	Multi-combine				40,	66,			65,		
	Others		30, 31,		09,						

Figure 3. Spectrum of boundaries across mandatory change.

Regulations issued by research funding agencies, research initiatives, academic institutions, and international scientific organizations further delineate responsibilities. While many documents specify restrictions on research data sharing, disciplines such as astronomy and geoscience tend to emphasize open data policies, whereas life sciences often impose stricter sharing constraints. These restrictions may encompass factors such as data scale, precision, timely, frequency, duration, spatial scope, data sources, themes, and multidimensional conditions.

In regulatory ambiguity areas, countries often issue guidelines and strategic documents to outline potential future measures and directions. A relevant example

is the CODATA publication, *Open Data in a Big Data World*, which states: *"Although it is tempting to suggest an embargo period, perhaps on the order of a year, it would be preferable for individual disciplines to develop procedures attuned to their specific needs, while avoiding undue delays."* This ambiguity is particularly evident in domains such as AI training datasets, cross-border data flows, and emerging technologies like quantum computing, where policy frameworks are still evolving. In February 2025, the OECD released a report titled *Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data*, highlighting that research institutions and universities frequently employ data scraping techniques for academic research and scientific inquiry. Although such activities are typically pursued for legitimate purposes, the use of international datasets may give rise to copyright and data privacy compliance challenges. For instance, scraped content used in studies on academic dissemination, social behavior, or public opinion trends may contain copyrighted materials—such as news articles, scholarly publications, or images—as well as personally identifiable information from sources like social media, user comments, and online forums. Cross-border scraping further raises the risk of triggering foreign data protection laws. Given the divergence in national copyright exceptions and the absence of a unified international framework, there is a growing expectation for the establishment of a coordinated governance mechanism for cross-border data scraping. If the proposed establishment of registration or transparency mechanisms for research-related data scraping were to be implemented, it could potentially reshape compliance requirements for research data in certain disciplinary fields.

Embargo period as a case of inconsistency detection

Standardizing and aligning quantitative constraints across legal and regulatory frameworks of varying levels and enforceability enable cross-national, cross-regional, and cross-institutional comparisons. This is crucial for identifying conflicts among these constraints, which pose significant challenges when research data is transferred across institutions, regions, or projects. Addressing such inconsistencies is one of the major operational difficulty researchers face in data-sharing practices.

Through the Prompt-based analysis of relevant policy documents, we identified a set of quantitative constraints. Among them, control over the embargo period is one of the most precisely quantified measures, with the embargo period itself serving as a key indicator. Figure 4 presents a schematic representation of embargo-related quantitative constraints, visually illustrating variations in start time, duration, and enforceability across different regulations. For instance, Document No. 02 mandates a two-year embargo period starting from the completion of data collection, whereas Document No. 28 recommends only one year. These discrepancies necessitate coordination and negotiation, as seen in the case of UK research funding agencies aligning embargo constraints when engaging in CODATA's international collaborations.

Similarly, quantitative constraints on data volume and frequency can be systematically mapped and compared, much like embargo periods. In contrast, subject-specific constraints on research data function more like controlled data catalogs, where sharing is restricted based on predefined classifications.

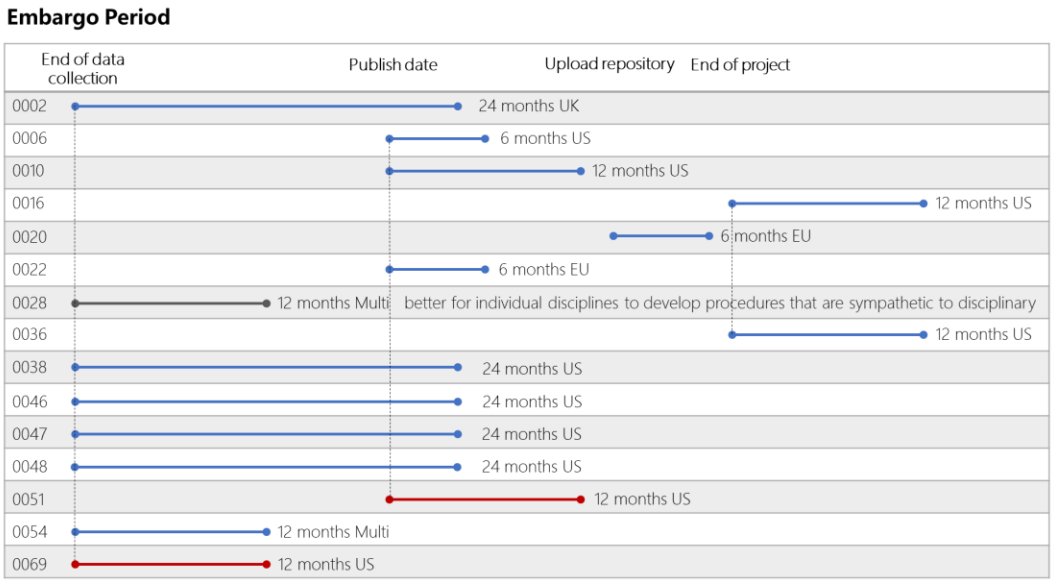


Figure 4. Schematic representation of quantitative boundaries for the embargo period indicator.

Preliminary remarks and limitations

Examining quantitative constraints in research data sharing policies provides a unified reference point for cross-domain collaboration, offering practical value for policy alignment. Our preliminary exploration has demonstrated that: (1) while some manual intervention and content review are still required, the prompt-based extraction method has proven successful and can be further refined into structured data. (2) Structured data effectively supports the visualization and mapping of quantitative constraints, enabling a more intuitive understanding of constraint variations, reducing the complexity of policy interpretation, and improving implementation efficiency.

Although this study highlights the importance of aligning research data policies across jurisdictions, achieving such coordination is fraught with legal and political complexity. From a legal perspective, civil law systems (e.g., the EU, Japan) rely on codified statutory exceptions, whereas common law systems (e.g., the United States) adopt interpretive doctrines such as fair use. Divergent views on data ownership, national sovereignty, and legal entitlements to access further complicate harmonization. Differences in regulatory culture, and institutional trust shape how jurisdictions approach research data governance. Even where overarching goals—such as advancing open science—are nominally shared, substantial asymmetries in enforcement capacity and legal infrastructure remain critical barriers to policy convergence.

Acknowledgments

This research was funded by the National Social Science Fund of China (Grant No. 22CTQ031).

References

Caufield, J. H., Hegde, H., Emonet, V., Harris, N. L., Joachimiak, M. P., Matentzoglu, N., Kim, H., Moxon, S., Reese, J. T., Haendel, M. A., Robinson, P. N., & Mungall, C. J. (2024). Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. *Bioinformatics*, *40*(3), btae104, Article btae104. https://doi.org/10.1093/bioinformatics/btae104

Chen, X., Li, Y., Fan, S. H., & Hou, M. S. (2024, Apr 26-28). *Judicial Text Relation Extraction Based on Prompt Tuning* 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition (CVIPPR), Xiamen, China. https://doi.org/10.1145/3663976.3664029

Durmaz, A. R., Thomas, A., Mishra, L., Murthy, R. N., & Straub, T. (2024). An ontology-based text mining dataset for extraction of process-structure-property entities. *Scientific Data*, *11*(1), 1112, Article 1112. https://doi.org/10.1038/s41597-024-03926-5

Li, Y. Z., Dong, L., Fan, X. X., Wei, R., Guo, S. J., Ma, W. Z., & Li, Z. X. (2024). New roles of research data infrastructure in research paradigm evolution. *Journal of Data and Information Science*, *9*(2), 104-119. https://doi.org/10.2478/jdis-2024-0011

Yang, Y. R., Chen, S. S., Zhu, Y. P., Liu, X. M., Ma, W., & Feng, L. (2024). Intelligent extraction of reservoir dispatching information integrating large language model and structured prompts. *Scientific Reports*, *14*(1), 14140, Article 14140. https://doi.org/10.1038/s41598-024-64954-0

Li, W. W., Leung, A. C. M., & Yue, W. T. (2023). Where is IT in information security? The interrelationship among IT investment, security awareness, and data breaches. *Mis Quarterly*, *47*(1), 317-342. https://doi.org/10.25300/misq/2022/15713

Amiri-Zarandi, M., Dara, R. A., Duncan, E., & Fraser, E. D. G. (2022). Big data privacy in smart farming: A review. *Sustainability*, *14*(15), 9120, Article 9120. https://doi.org/10.3390/su14159120

Cao, L. J., & Zhang, Z. Q. (2022). Research Progress on Quantitative Methods of Policy Texts from the Perspective of Policy Informatics. *Library and Information*(6), 70-82.

Li, Y., Liu, X., Li, Z., Yin, X., & Wu, M. (2022). Study on conceptual analysis model of scientific data security boundary: from the perspective of stakeholders. *Bulletin of National Natural Science Foundation of China*, *36*(2), 339-347.

Majeed, A. (2021). Towards privacy paradigm shift due to the pandemic: a brief perspective. *Inventions*, *6*(2), 24, Article 24. https://doi.org/10.3390/inventions6020024

Xiang, D., & Cai, W. (2021). Privacy protection and secondary use of health data: strategies and methods. *Biomed Research International*, *2021*, Article ID 6967166, Article 6967166. https://doi.org/10.1155/2021/6967166

Ducato, R. (2020). Data protection, scientific research, and the role of information *Computer Law and Security Review*, *37*, 105412.

Yang, C., Huang, C., & Su, J. (2020). A bibliometrics-based research framework for exploring policy evolution: A case study of China's information technology policies. *Technological Forecasting and Social Change*, *157*, 120116. https://doi.org/https://doi.org/10.1016/j.techfore.2020.120116

Zigomitros, A., Casino, F., Solanas, A., & Patsakis, C. (2020). A Survey on Privacy Properties for Data Publishing of Relational Data. *Ieee Access*, *8*, 51071-51099. https://doi.org/10.1109/Access.2020.2980235

Bardach, E., & Patashnik, E. M. (2019). *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*. SAGE Publications. https://books.google.com.hk/books?id=cil_DwAAQBAJ

George, A. M. (2019). The National Security Implications of Cyberbiosecurity. *Frontiers in Bioengineering and Biotechnology*, *7*, 51-54, Article 51. https://doi.org/10.3389/fbioe.2019.00051

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 254-277. https://doi.org/10.1093/pan/mpu019

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267-297. https://doi.org/10.1093/pan/mps028

Supplementary Information

Table SI-1. List of 72 Documents Related to Scientific Data Sharing and Management.

OID	File Name	Countries	Type	Subtype	Issuing Date	Enforceability	Access Address
01	ESRC data citation: what you need to know	UK	Rules	Rules_Sponsor_Public	2012		https://www.ukri.org/publications/data-citation-what-you-need-to-know/
02	NERC Data Policy	UK	Rules	Rules_Sponsor_Public	2019		https://www.ukri.org/publications/nerc-policies/
03	STFC scientific data policy	UK	Rules	Rules_Sponsor_Public	2019		https://www.ukri.org/publications/stfc-scientific-data-policy/
04	Guidance on best practice in the management of research data	UK	Rules	Rules_Government	2018		https://www.ukri.org/publications/guidance-on-best-practice-in-the-management-of-research-data/
05	Data protection policy	UK	Rules	Rules_Government	2022	Y	https://www.ukri.org/publications/data-protection-policy/
06	Open access policy	UK	Rules	Rules_Sponsor_Private	2025		https://wellcome.org/grant-funding/guidance/open-access-guidance/open-access-policy
07	Data sharing and management policy	UK	Rules	Rules_Sponsor_Public	2022		https://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/data-sharing-and-management-policy
08	SRS Research and Data Access Policy	UK	Rules	Rules_Government	2023		https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/onsresearchanddataaccesspolicy
09	Data sharing guidance for researchers seeking permission for secure access to data	UK	Guidance	Guidance	2022		https://assets.publishing.service.gov.uk/media/62038afa8fa8f510b357cc44/data-sharing-guidance-researchers.pdf
10	Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation	US	Rules	Rules_Sponsor_Public	2015		https://new.nsf.gov/reports/performance/public-access-plan-todays-data-tomorrows-discoveries
11	Data Management and Sharing Plan Guidelines (in PAPPG II.D.2(ii)) Proposal & Award Policies & Procedures Guide (PAPPG) (NSF 24-1) Chapter II: Proposal Preparation Instructions	US	Rules	Rules_Sponsor_Public	2024		https://new.nsf.gov/policies/pappg/24-1/ch-2-proposal-preparation#ch2D2i-ii
12	NASA's Public Access Plan	US	Rules	Rules_Sponsor_Public	2023		https://researchdata.wvu.edu/regulations-and-policies/public-access-and-dms-policies/nasa-s-public-access-plan
13	USDA Public Access and Open Science Plan	US	Rules	Rules_Sponsor_Public	2023		https://researchdata.wvu.edu/regulations-and-policies/public-access-and-dms-policies/usda-public-access-and-open-science-plan
14	DOE Public Access Plan	US	Rules	Rules_Sponsor_Public	2023		https://researchdata.wvu.edu/regulations-and-policies/public-access-and-dms-policies/doe-public-access-plan
15	DOE Policy for Digital Research Data Management: Glossary	US	Rules	Rules_Sponsor_Public	2015		https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management-glossary#Data%20Sharing
16	Data Policy and Guidance	US	Rules	Rules_Sponsor_Public	2018		https://www.usgs.gov/media/files/casc-data-sharing-policy

OID	File Name	Countries	Type	Subtype	Issuing Date	Enforceability	Access Address
17	Public Access to Results of Federally Funded Research at the U.S. Geological Survey: Scholarly Publications and Digital Data (ver. 2.0)	US	Rules	Rules_Sponsor_Public	2023		https://www.usgs.gov/media/files/public-access-results-federally-funded-research-us-geological-survey-scholarly
18	JRC Data Policy	EU	Rules	Rules_Sponsor_Public	2019		https://publications.jrc.ec.europa.eu/repository/handle/JRC115832
19	Guidelines on FAIR Data Management in Horizon 2020	EU	Rules	Rules_Project	2016		https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
20	Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020	EU	Rules	Rules_Project	2017		https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
21	Open Research Data and Data Management Plans	EU	Rules	Rules_Government	2022	Y	https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf
22	Guidelines on the Implementation of Open Access to Scientific Publications and Research Data in Projects supported by the European Research Council under Horizon 2020	EU	Guidance	Guidance	2016		https://erc.europa.eu/sites/default/files/ERC_Guidelines_Implementation_Open_Access.pdf
23	General Data Protection Regulation, GDPR	EU	Regulations	Regulations	2016	Y	https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
24	Data Act	EU	Act	Act	2023	Y	https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng
25	Data Governance Act, DGA	EU	Act	Act	2022	Y	https://eur-lex.europa.eu/eli/reg/2022/868/oj/eng
26	OECD Principles and Guidelines for Access to Research Data from Public Funding	Multi	Rules	Rules_International Organization	2007		https://www.oecd-ilibrary.org/science-and-technology/oecd-principles-and-guidelines-for-access-to-research-data-from-public-funding_9789264034020-en-fr
27	CODATA Strategic Plan 2015	Multi	Strategy	Strategy_International Organization	2015		https://zenodo.org/record/165830#.XusKixbiuM8
28	Open data in a big data world	Multi	Strategy	Strategy_International Organization	2015		https://council.science/wp-content/uploads/2017/04/open-data-in-big-data-world_long.pdf
29	ICSU-WDS Bylaws	Multi	Rules	Rules_International Organization	2023		https://worlddatasystem.org/wp-content/uploads/2023/05/WDS_bylaws_19April2023.pdf
30	Export Administration Regulations	US	Regulations	Regulations	2024	Y	https://media.bis.gov/regulations/ear
31	EU Regulation on Export Controls for Dual-Use Items	EU	Regulations	Regulations	2021	Y	https://eur-lex.europa.eu/eli/reg/2021/821/oj/eng
32	Open Data Directive	EU	Directive	Directive	2019	Y	https://eur-lex.europa.eu/eli/dir/2019/1024/oj
33	A framework for the free flow of non-personal data in the European Union	EU	Regulations	Regulations	2018	Y	https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32018R1807
34	WMO Unified Data Policy	Multi	Rules	Rules_International Organization	2022		https://library.wmo.int/viewer/58009/download?file=WMO_Unified_Data_Policy_brochure_en.pdf&type=pdf&navigator=1

OID	File Name	Countries	Type	Subtype	Issuing Date	Enforceability	Access Address
35	Rules Governing the Distribution and Dissemination of ECMWF Real-Time Products	Multi	Rules	Rules_International Organization	1994		https://www.ecmwf.int/sites/default/files/Rules_real_time_products.pdf
36	EOL Data Policy	US	Rules	Rules_Institution	2014		https://www.eol.ucar.edu/content/eol-data-policy
37	GEO Data Management and Sharing Plan Guidance	US	Rules	Rules_Sponsor_Public	2024		https://new.nsf.gov/geo/data-management-sharing-plans
38	Update to the Division of Earth Sciences (EAR) Data and Sample Policy	US	Rules	Rules_Sponsor_Public	2023		https://www.nsf.gov/pubs/2023/nsf23131/nsf23131.jsp
39	Data Policy for the IGBP	Multi	Rules	Rules_International Project	1994		https://pastglobalchanges.org/sites/default/files/download/docs/IGBP_Data_Policy.pdf
40	Ensuring Free, Immediate, and Equitable Access to Federally Funded Research	US	Rules	Rules_National	2022		https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf
41	The FAIR Guiding Principles for scientific data management and stewardship	Multi	Guidance	Guidance	2016		https://www.nature.com/articles/sdata201618
42	The CARE Principles for Indigenous Data Governance	Multi	Guidance	Guidance	2020		https://datascience.codata.org/articles/10.5334/dsj-2020-043
43	A Vision for NSF Earth Sciences 2020-2030	US	Strategy	Strategy_Institution	2020		https://nap.nationalacademies.org/catalog/25761/a-vision-for-nsf-earth-sciences-2020-2030-earth-in
44	NSF Public Access Plan 2.0 Ensuring Open, Immediate and Equitable Access to National Science Foundation Funded Research	US	Rules	Rules_Sponsor_Public	2023		https://nsf-gov-resources.nsf.gov/pubs/2023/nsf23104/nsf23104.pdf
45	Desirable Characteristics of Data Repositories for Federally Funded Research	US	Rules	Rules_National	2022		https://repository.si.edu/bitstream/handle/10088/113528/Desirable%20Characteristics%20of%20Data%20Repositories.pdf?sequence=3&isAllowed=y
46	Division of Ocean Sciences (OCE) Sample and Data Policy	US	Rules	Rules_Sponsor_Public	2024		https://www.nsf.gov/pubs/2024/nsf24124/nsf24124.jsp
47	Division of Ocean Sciences (OCE) Sample and Data Policy	US	Rules	Rules_Sponsor_Public	2016		https://www.nsf.gov/pubs/2017/nsf17037/nsf17037.jsp
48	Office of Polar Programs Data, Code, and Sample Management Policy	US	Rules	Rules_Sponsor_Public	2022		https://new.nsf.gov/funding/information/dcl-office-polar-programs-data-code-sample-management-policy
49	Proprietary and Sensitive Data	US	Rules	Rules_Sponsor_Public	2024		https://www.usgs.gov/data-management/proprietary-and-sensitive-data
50	Survey Manual 502.5 - Fundamental Science Practices: Safeguarding Unpublished USGS Scientific Information and Associated Materials	US	Rules	Rules_Sponsor_Public	2019		https://www.usgs.gov/survey-manual/5025-fundamental-science-practices-safeguarding-unpublished-usgs-scientific
51	Increasing Access to the Results of Federally Funded Scientific Research,	US	Directive	Directive	2013	Y	https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
52	Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure	US	Rules	Rules_Sponsor_Public	2003		https://www.fgdc.gov/policyandplanning/executive_order
53	National Geospatial Data Asset Management Plan	US	Rules	Rules_Sponsor_Public	2014		https://www.fgdc.gov/policyandplanning/a-16/ngda-management-plan
54	IODP Sample, Data, and Obligations Policy and Implementation	Multi	Rules	Rules_International Project	2018		https://www.iodp.org/top-resources/program-documents/policies-and-guidelines/519-iodp-sample-data-and-obligations-policy-

OID	File Name	Countries	Type	Subtype	Issuing Date	Enforceability	Access Address
							implementation-guidelines-may-2018-for-expeditions-starting-october-2018-and-later/file
55	Guide to Best Practices for Generalising Sensitive Species Occurrence data	Multi	Rules	Rules_International Organization	2023		https://assets.ctfassets.net/uo17ejk9rkwj/6ie7n89wYMA6IcGKyoqW2/46d527fcd192ac18ec6c0be909bb8f20/gbif_Sensitive_Data_guide_en_v1.pd
56	1100.2 - Editorial Review of U.S. Geological Survey Publication Series Information Products	US	Rules	Rules_Sponsor_Public	2021		https://www.usgs.gov/survey-manual/11002-editorial-review-us-geological-survey-publication-series-information-products
57	NAO 212-15B: Management of NOAA Data and Information	US	Rules	Rules_Sponsor_Public	2023		https://www.noaa.gov/organization/administration/nao-212-15-Management-of-NOAA-Data-and-Information
58	Land Remote Sensing Policy Act	US	Act	Act	1992	Y	https://www.congress.gov/bill/102nd-congress/house-bill/6133
59	Management of NOAA Data and Information Data Management Handbook	US	Rules	Rules_Sponsor_Public	2024		https://nosc.noaa.gov/EDMC/documents/NAO_212-15B-Data_Mgt_Handbook-2024-Oct-1_remediated.pdf
60	National Space Policy of the United States	US	Strategy	Strategy_National	2020		https://trumpwhitehouse.archives.gov/wp-content/uploads/2020/12/National-Space-Policy.pdf
61	American Space Commerce Free Enterprise Act of 2018	US	Act	Act	2018	Y	https://www.govinfo.gov/content/pkg/BILLS-115hr2809rfs/pdf/BILLS-115hr2809rfs.pdf
62	Controlled Unclassified Information (CUI) Procedure	US	Directive	Directive	2024	Y	https://www.epa.gov/system/files/documents/2024-07/controlled_unclassified_information_procedure.pdf
63	NASA’S PUBLIC ACCESS PLAN Increasing Access to the Results of Scientific Research	US	Rules	Rules_Sponsor_Public	2023		https://www.nasa.gov/wp-content/uploads/2021/12/nasa-ocs-public-access-plan-may-2023.pdf
64	ESA Data Policy for ERS, Envisat and Earth Explorer missions	EU	Rules	Rules_Project	2012		https://earth.esa.int/eogateway/documents/20142/1564626/ESA-Data-Policy-ESA-PB-EO-2010-54.pdf
65	Study on the COPERNICUS Data Policy POST-2020	Multi	Rules	Rules_International Project	2019		https://data.europa.eu/en/news-events/news/study-copernicus-data-policy-post-2020
66	Updated ESA Earth Observation Data Policy	EU	Rules	Rules_Sponsor_Public	2023		https://earth.esa.int/eogateway/documents/d/earth-online/esa-eo-data-policy
67	Public Access Plan	US	Rules	Rules_Sponsor_Public	2023		https://www.energy.gov/sites/default/files/2023-07/DOE%20Public%20Access%20Plan%202023%20-%20Final.pdf
68	Federal Data Strategy Data Ethics Framework	US	Rules	Rules_Government	2020	Y	https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf
69	Freedom of Information Act	US	Act	Act	2016	Y	https://www.congress.gov/114/plaws/publ185/PLAW-114publ185.pdf
70	Revise Freedom of Information Act	US	Act	Act	2022	Y	https://www.justice.gov/oip/freedom-information-act-5-usc-552
71	DOE Requirements and Guidance for Digital Research Data Management	US	Rules	Rules_Sponsor_Public	2024		https://www.energy.gov/datamanagement/doe-requirements-and-guidance-digital-research-data-management
72	Data Ethics Framework	UK	Rules	Rules_Government	2020	Y	https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020

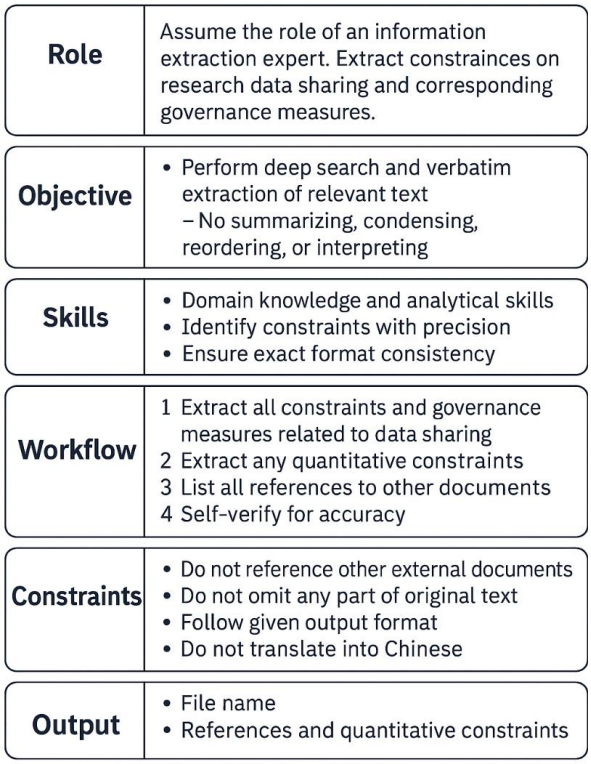


Figure SI-1. Prompt Framework: Role-Objective-Skill-Workflow-Constraint-Output.

https://doi.org/10.51408/issi2025_129

Mapping the Social Structure of Philosophy of Science Through Large-Scale Acknowledgments Analysis

Eugenio Petrovich[1], Edoardo Fazzini[2], Lorenzo Gandolfi[3]

[1] *eugenio.petrovich@unito.it*
University of Turin, Dept. of Philosophy and Education Sciences, Via Sant'Ottavio 20, Turin (Italy)

[2] *edoardo.fazzini@unibe.ch*
University of Bern, Institute of Philosophy, Länggassstrasse 49a, 3012 Bern (Switzerland)

[3] *lorenzo.gandolfi@iusspavia.it*
IUSS Pavia, Palazzo del Broletto, Piazza della Vittoria 15, 27100 Pavia PV (Italy)

Abstract

The acknowledgments in scientific publications offer a unique perspective on the complex web of socio-cognitive relationships underlying the production of knowledge. Acknowledgment analysis enables us to highlight the role of funding institutions, reconstruct informal collaboration patterns invisible to co-authorship analysis, and measure a distinct form of prestige beyond authorships and citations. This study leverages acknowledgment analysis to investigate the fine-grained social structure of philosophy of science, a research field whose social dimension has thus far remained unexplored. Specifically, we aim to: 1) identify the scholars who receive the most acknowledgments in the field and examine their roles in professional associations; 2) analyze how acknowledgments are distributed across the community and the factors affecting the number of mentions received; and 3) map the social communities within philosophy of science, exploring whether they are organized around topics, methodological approaches, or professional associations. Our findings reveal that acknowledgments are prevalent in philosophy of science, with 79% of articles including them, and that the average acknowledgment mentions 5.3 individuals—significantly higher than the average number of co-authors per article (1.3). Most acknowledged individuals are prominent philosophers of science who play key roles in professional associations. In terms of distribution, mentions are highly concentrated among a few individuals, with the top 10% of acknowledged scholars receiving nearly half of all mentions. Mentions are most strongly predicted by academic awards, productivity in philosophy of science journals, leadership roles in professional associations, and affiliation with English-speaking institutions, with smaller effects for gender and general publication metrics. Finally, the co-acknowledgment network shows that clusters of frequently co-mentioned individuals are organized around both topics (e.g., philosophies of specific sciences) and methodological approaches (e.g., formal vs. historical philosophy of science).

Introduction

In scientometrics, *acknowledgement analysis* is a relatively underdeveloped field of investigation, especially when compared to well-established domains such as citation analysis or publication analysis.

Still, the acknowledgments of academic publications are a rich source of information, which are able to illuminate streams of research funding (Huang & Huang, 2018), informal collaboration patterns (Cronin et al., 2003; Petrovich, 2021), and prestige dynamics within research fields (Costas & Leeuwen, 2012; Petrovich, 2024). The acknowledgments, in fact, offer an unique perspective on the complex web of *socio-cognitive relationships* that underlie research, highlighting actors and

social processes that remain invisible when the analysis is confined to standard data considered in scientometrics, such as authorships and citations (Cronin et al., 2004). The acknowledgements are especially valuable to investigate *scientific collaboration* in those fields, such as the social sciences and humanities, in which formal co-authorship is less common. Including the persons mentioned in the acknowledgements in addition to formal co-authors reveals, in fact, that the intensity of collaboration in these fields is not different from what is observed in the natural sciences (Paul-Hus et al., 2017). Similarly, acknowledgments networks are able to reveal portions of the social structure of research fields that remain invisible to standard co-authorship networks (Petrovich, 2022).

Moreover, being mentioned in an acknowledgment is a sign of *prestige*: persons that are frequently thanked in the acknowledgments are in fact prominent figures in their respective fields (Cronin, 1995). Among the most mentioned acknowledgees in economics, for instance, there are the editors of the most prestigious economics journals, as well as winners of important economics prizes, including several Nobel laureates (Baccini & Petrovich, 2022). Similarly, the most mentioned figures in biology are respected mentors and recognized experts of specific organisms (McCain, 2024). In this sense, the acknowledgments constitute the third angle of the "reward triangle" of science, along with authorships and citations, and offer a further way to measure prestige in a research community (Costas & Leeuwen, 2012).

The aim of this paper is to provide the first large-scale quantitative analysis of the acknowledgments in contemporary *philosophy of science* in order to shed light on the social structure of this research field.

Philosophy of science is an interesting case study for a four-fold reason. First, it is one of the few areas of philosophy that has been quite extensively investigated through quantitative methods, both via text mining (Malaterre et al., 2021) and citation analysis (Khelfaoui et al., 2021; McLevey et al., 2018; Petrovich & Viola, 2024). Still, a detailed study of the social dimension of the field via acknowledgments analysis is lacking in the literature. Second, acknowledgments data can provide a better estimate of the intensity of scientific collaboration in philosophy of science, a field where co-authorship is relatively uncommon (e.g., only 17% of recent publications in the journal *Philosophy of Science* are multi-authored). Third, philosophy of science has greatly diversified in the last two decades, both from the point of view of social structures and methodological orientations. The Philosophy of Science Association (PSA), founded back in 1933, has been for long time the only professional association of philosophers of science. From 1990s, however, new associations have been established: the International Society for the History of Philosophy of Science (HOPOS) in 1994, the European Philosophy of Science Association (EPSA) in 2005, the Committee for Integrated HPS (&HPS), the Society for Philosophy of Science in Practice (SPSP), both in 2006, and the Consortium for Socially Relevant Philosophy of/in Science and Engineering (SRPoiSE) in 2014, not to mention the numerous associations devoted to the philosophy of specific sciences. All these associations aim to promote and advance philosophy of science, but with slightly different methodological orientations: for instance, HOPOS promotes historical research on the philosophy of science widely

understood, while PSPS promotes the epistemological analysis of scientific practices. Mapping the social landscape of the field can allow us to better understand the impact of these newer associations on the development of the field. Lastly, philosophy of science is itself divided into specialties: along with general philosophy of science, there are philosophies of the different sciences, such as philosophy of physics, mathematics, economics, neuroscience, and so on. To investigate the sociology of the field, it is crucial to understand whether its social structure reflects this specialization, i.e., whether intellectual specialties correspond to social sub-communities, or not.

In the light of these interesting characteristics of philosophy of science, the present study aims specifically to answer the following research questions:

R1) Who are the scholars who receive most mentions in philosophy of science? Do they play key roles in old and new professional associations of the field?

R2) How is prestige distributed in the community? Is it concentrated in few individuals or equally shared among social actors? Is it equally distributed between genders? What factors influence the accumulation of prestige?

R3) What are the social communities in which philosophers of science are divided? Are communities organized around topics, methodological approaches, professional associations?

Data and Methods

Following previous quantitative studies on philosophy of science (Malaterre et al., 2021), we operationally defined the field based on a widely accepted list of leading disciplinary journals. In particular, we focused on the following 8 journals: *Erkenntnis, Philosophy of Science, Synthese, Studies in History and Philosophy of Science, British Journal for the Philosophy of Science, Foundations of Science, Journal for General Philosophy of Science, and European Journal for Philosophy of Science*. Of the 8,327 publications appeared in these journals between 2005 and 2019, we retained research articles ($n = 6{,}826$), leaving aside book reviews, commentaries, editorials, and other minor document types. This set of articles corresponds to our bibliometric delineation of philosophy of science.

All metadata of these articles, including authorship and cited references, were downloaded from Web of Science. The acknowledgments appearing in the articles, on the other hand, were manually collected from the articles' electronic version and attached to the main dataset. We focused only on acknowledgments that were clearly recognizable as such, ignoring minor acknowledgements appearing in the main text and in standard footnotes.

To extract the names of the persons thanked in the acknowledgments (henceforth, the *acknowledgees*), we used Named-Entity Recognition, a Natural Language Processing technique that is able to identify and classify into pre-defined categories named entities occurring in pieces of natural language. Specifically, we used the NER module of the Python library *spaCy* (https://spacy.io) to extract from our corpus of acknowledgement texts around 49,000 mentions to around 20,000 distinct named entities.

The raw output of the NER was then manually cleaned and consolidated. First, misclassifications were manually corrected: false positives, i.e., entities wrongly classified as PERSON ($n = 781$), were excluded from the list of acknowledgees, while false negatives, i.e., entities that were not classified as PERSON even if they were persons ($n = 298$), were added. Second, name variants were identified and consolidated. Due to the informal nature of acknowledgments, diminutives (e.g., "Jon Kvanvig") are often used alongside full names (e.g., "Jonathan Kvanvig"), leading to multiple ways of referring to the same individual. To ensure accurate mention statistics, these variants needed to be standardized. The identification process combined custom Python scripts based on string similarity and fuzzy matching with manual inspection and validation. After this consolidation, we reduced the initial 10,570 entities classified as persons in the raw output to a refined list of 9,029 distinct acknowledgees (-15%).

Authorship data were similarly consolidated, as Web of Science does not provide unique identifiers for authors. Of the 4,835 raw distinct author strings, we build a list of 4,395 standardized authors (-9%). The lists of acknowledgees and authors were finally merged, obtaining a list of 10,980 actors, i.e., persons that appeared as authors and/or acknowledgees in our dataset. This merging allowed us to remove few false self-mentions that occurred when the name of the authors of an article appeared in the acknowledgement as well, for instance when the authors acknowledged some funding body.

Affiliation data with philosophy of science associations were manually collected from their respective websites, focusing on PSA, EPSA, &HPS, SPSP, SRPoiSE, and SMS (Society for the Metaphysics of Science). We focused only on members of governing bodies (e.g., presidents and officials), leaving aside simple membership. Similarly, the names of recipients of the most prestigious philosophy of science prizes (Lakatos Award, Popper Prize, and the Hempel Award) were retrieved from online sources.

Academic affiliations—including institution name, city, and country—were assigned to actors by retrieving their Scopus profiles through an automated search using the Scopus API. To account for academic mobility, the affiliation corresponding to the year of the authorship or acknowledgment was used. 81% of the mentions were successfully linked to an affiliation. Bibliometric statistics of actors (citation counts and publication counts) were also retrieved from Scopus profiles. Lastly, gender was assigned using the *genderize.io* service, based on the actor's first name and primary country of affiliation, to account for names (such as "Andrea") that vary in gender across different countries.

To answer R1 and R2, we developed an indicator M_a that measures the prestige of an actor a in the community, based on the number of distinct acknowledgments in which a is mentioned. The higher the value of M_a, the greater the prestige of a in the community. To answer R3, we constructed an Acknowledgees Co-Mention Network (ACM) using the techniques developed by Petrovich (2022) to map the social structure of research fields. In the ACM, acknowledgees are connected when they are co-mentioned in the same acknowledgement, with the strength of the links equal to the number of acknowledgements in which they are mentioned together.

Clusters of densely interconnected acknowledgees in the ACM correspond to social communities within a research field (Petrovich, 2022).

Results and Discussion

79% ($n = 5{,}376$) of the articles in our dataset included acknowledgments, with the percentage increasing linearly over time from 74% in 2005 to 84% in 2019. The average acknowledgment is 60 words long (st. dev. = 40, median = 52, min = 4, max = 391) and 87% of the acknowledgments ($n = 4{,}660$) mentioned at least one acknowledgee. Considering only this subset, the average number of mentioned acknowledgees per article is 5.3 (st. dev. = 4.4, median = 4, min = 1, max = 44), with some variance across journals. Note that the average number of mentioned acknowledgees per article is significantly higher than the average number of co-authors per article: 5.3 against 1.3. This difference shows that co-authorship severely underestimates the rate of collaboration in philosophy of science.

2,444 actors appear both as authors and as acknowledgees in our dataset, meaning that 55.8% of authors are mentioned also in the acknowledgments, and 27% of acknowledgees write also an article. The high number of acknowledgees that do not appear as authors ($n = 6{,}595$) shows that the population of actors contributing to the development of philosophy of science extends significantly beyond that of formal authors.

To address R1, we constructed the ranking of acknowledgees based on the M indicator. Table 1 shows the acknowledgees with the top-10 highest M.

Table 1. Top-10 most-mentioned acknowledgees in philosophy of science (*=President).

Rank	*Actor*	*Mentions*	*Articles*	*Association(s)*
1	Elliott Sober	104	7	PSA*
2	John Norton	97	15	HPS, PSA
3	Carl Craver	66	7	
4	Anjan Chakravartty	65	8	SRPoiSE, PSA
5	Stephan Hartmann	62	14	EPSA*, PSA
6	David Chalmers	61	0	
7	Alan Hajek	60	7	PSA
8	James Woodward	57	11	PSA*
9	Branden Fitelson	55	8	
10	Hannes Leitgeb	54	8	

The most-mentioned persons in philosophy of science are all academic philosophers playing prominent roles in the profession. Elliott Sober (rank 1) and James Woodward (rank 8) have both served as President of the PSA in 2003-2004 and 2011-2012, while Stephan Hartmann (rank 5) has been President of EPSA in 2013-2015. SPSP, HOPOS, and SMS are not represented in the ranking, suggesting that their officials play a less relevant role in the social landscape of philosophy of science. Interestingly, among the most mentioned philosophers we find also David Chalmers, who is a prominent analytic philosopher rather than a philosopher of

science *strictu sensu* (note that he does not author any article in the corpus). Moreover, all top-mentioned are male. The first woman, Nancy Cartwright is in rank 11, and among the 140 acknowledgees with more than 20 mentions, there are only 10 women (7%).

To address R2, we analysed the distribution of mentions across the entire population of acknowledgees and by gender. In our dataset, a total of 24,912 mentions are distributed among 9,029 distinct acknowledgees. The average number of mentions per acknowledgee is 2.7 (median = 1, standard deviation = 5, minimum = 1, maximum = 104), but the Gini coefficient of 0.53 indicates significant inequality in the distribution of mentions. Specifically, 80% of acknowledgees collect only 37% of all mentions, while the top 10% most-mentioned acknowledgees collect nearly 50%. This skewed distribution is typical of scientometric variables, demonstrating that the form of prestige captured by acknowledgments is, in fact, concentrated among a small number of individuals, similarly to what is observed with citations and authorships. In terms of gender, the overall population of actors is characterized by a significant disparity, with only 22% of women. Note that the proportion of women in the authors (80.4%) is slightly lower than the proportion of women in the acknowledgees (77.9%), suggesting a possible bias against women in accessing formal authorship.

A multiple regression analysis was conducted to examine the predictors of mentions. The number of awards was the strongest predictor, with each additional prize associated with 10.1 more mentions (SE = 0.57, $p < 0.001$). The number of publications in philosophy of science journals, the number of governing roles in professional associations, and affiliation with English-speaking countries also had substantial effects, increasing mentions by 1.56 (SE = 0.034), 1.12 (SE = 0.12), and 1.38 (SE = 0.13), respectively (all $p < 0.001$). Gender showed a smaller but statistically significant effect, with men receiving roughly 0.54 more mentions than women (SE = 0.16, $p < 0.001$). Overall publication count had a small but significant positive effect ($\beta = 0.0060$, SE = 0.0016, $p < 0.001$), while citation count showed a small negative association ($\beta = –0.000044$, SE = 0.000016, $p = 0.006$). The model explains approximately 40.5% of the variance in mention counts (adjusted $R^2 = 0.405$), indicating that academic recognition—particularly in the form of prizes, productivity in philosophy of science, and institutional visibility through professional associations—is a major driver of acknowledgment practices.

Lastly, to address R3, we constructed the Acknowledgees Co-Mention Network including all the acknowledgees receiving at least 10 mentions ($n = 447$). An interactive visualization of the network, created with VOSviewer (van Eck & Waltman, 2010) and supplemented with further statistics and information, is available on-line at https://tinyurl.com/288wkbxt. VOSviewer clustering algorithm identifies 6 clusters at resolution 1, which can be straightforwardly labelled based on the specialization of the acknowledgees they include. Interestingly, cluster 1 includes mainly analytic philosophers rather than philosophers of science; cluster 2 includes philosophers of science working on general philosophy of science, frequently using formal methods such as probability theory; cluster 3 philosophers of science working on integrated history and philosophy of science; cluster 4 philosophers of physics;

cluster 5 philosophers of life sciences; cluster 6 philosophers of mind sciences. Clusters appear therefore to be organized both around specialties (philosophies of specific sciences) and methodological approaches (formal vs historically-informed philosophy of science). Note that, in the overall network, the philosophy of physics cluster is the most isolated, showing that philosophers of physics constitute a tight sub-community with relatively few connections with the rest of the field. Philosophy of physics is also the cluster with the highest average number of mentions as well as the one which includes the highest number of awardees (see Table 2). The cluster with the highest number of officials, however, is integrated history and philosophy of science, with 32 officials (43% of the cluster's members are official in at least one of the associations covered). This cluster is also the one with the highest proportion of women (25.3%). Officials of the PSA, the oldest association, can be found in all clusters, showing the influence of the association on the entire field. Officials of the younger EPSA, on the other hand, are mainly concentrated in the general philosophy of science cluster, while, unsurprisingly, officials of historically- and practice-oriented associations (SPSP, HOPOS, &HPS) can be found mainly in the integrated history and philosophy of science cluster.

Table 2. Cluster-level statistics of the philosophy of science ACM network.

Cluster	*Label*	*Members*	*Avg. Mentions*	*Awards*	*Associations officials (%)*	*Women prop.(%)*
1	Analytic philosophy	117	16.7	0	1 (0.9%)	11.2
2	General philosophy of science	94	19.5	5	20 (21%)	10.6
3	Integrated Hist. & Phil. of Science	75	19.0	7	32 (43%)	25.3
4	Philosophy of physics	73	22.2	13	20 (12%)	16.4
5	Philosophy of biology	58	22.0	9	7 (14%)	10.3
6	Philosophy of mind sciences	29	20.7	1	4 (27%)	10.3

Conclusions and next steps in the research

This preliminary investigation of philosophy of science via acknowledgment analysis has shown that the acknowledgments of academic articles offer precious insights on the social structure of this research field. Our data has allowed us to identify prominent figures in the field (R1), determine how prestige is distributed in the community and the factors governing it (R2), and map the communities in which philosophers of science are divided (R3), highlighting in particular the role that different professional associations play in the field.

The next step in the research is to extend the analysis to the institutional level, in order to identify the most prominent research centers in philosophy of science and to examine the role of homophily (i.e., similarity in characteristics) in shaping the relationships between authors and acknowledgees.

References

Baccini, A., & Petrovich, E. (2022). Normative versus strategic accounts of acknowledgment data: The case of the top-five journals of economics. *Scientometrics, 127*(1), 603–635. https://doi.org/10.1007/s11192-021-04185-6

Costas, R., & Leeuwen, T. N. (2012). Approaching the "reward triangle": General analysis of the presence of funding acknowledgments and "peer interactive communication" in scientific publications. *Journal of the American Society for Information Science and Technology, 63*(8), 1647–1661. https://doi.org/10.1002/asi.22692

Cronin, B. (1995). *The Scholar's Courtesy: The Role of Acknowledgement in the Primary Communication Process*. Taylor Graham.

Cronin, B., Shaw, D., & Barre, K. L. (2004). Visible, less visible, and invisible work: Patterns of collaboration in 20th century chemistry. *Journal of the American Society for Information Science and Technology, 55*(2), 160–168. https://doi.org/10.1002/asi.10353

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54*(9), 855–871. https://doi.org/10.1002/asi.10278

Huang, M.-H., & Huang, M.-J. (2018). An analysis of global research funding from subject field and funding agencies perspectives in the G9 countries. *Scientometrics, 115*(2), 833–847. https://doi.org/10.1007/s11192-018-2677-y

Khelfaoui, M., Gingras, Y., Lemoine, M., & Pradeu, T. (2021). The visibility of philosophy of science in the sciences, 1980–2018. *Synthese*. https://doi.org/10.1007/s11229-021-03067-x

Malaterre, C., Lareau, F., Pulizzotto, D., & St-Onge, J. (2021). Eight journals over eight decades: A computational topic-modeling approach to contemporary philosophy of science. *Synthese, 199*(1–2), 2883–2923. https://doi.org/10.1007/s11229-020-02915-6

McCain, K. W. (2024). Collaboration at the phylum level: Coauthorship and acknowledgment patterns in the world of the water bears (phylum Tardigrada). *Scientometrics, 129*(10), 6089–6125. https://doi.org/10.1007/s11192-024-05036-w

McLevey, J., Graham, A. V., McIlroy-Young, R., Browne, P., & Plaisance, K. S. (2018). Interdisciplinarity and insularity in the diffusion of knowledge: An analysis of disciplinary boundaries between philosophy of science and the sciences. *Scientometrics, 117*(1), 331–349. https://doi.org/10.1007/s11192-018-2866-8

Paul-Hus, A., Mongeon, P., Sainte-Marie, M., & Larivière, V. (2017). The sum of it all: Revealing collaboration patterns by combining authorship and acknowledgements. *Journal of Informetrics, 11*(1), 80–87. https://doi.org/10.1016/j.joi.2016.11.005

Petrovich, E. (2021). Acknowledgments. Informal collaboration and symbolic power in recent analytic philosophy. *Logique et Analyse, 256*, 425–448. https://doi.org/10.2143/LEA.256.0.3290352

Petrovich, E. (2022). Acknowledgments-based networks for mapping the social structure of research fields. A case study on recent analytic philosophy. *Synthese, 200*(3), 204. https://doi.org/10.1007/s11229-022-03515-2

Petrovich, E. (2024). *A Quantitative Portrait of Analytic Philosophy: Looking Through the Margins*. Springer International Publishing. https://doi.org/10.1007/978-3-031-53200-9

Petrovich, E., & Viola, M. (2024). Mapping the philosophy and neuroscience nexus through citation analysis. *European Journal for Philosophy of Science*, *14*(4), 60. https://doi.org/10.1007/s13194-024-00621-5

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

https://doi.org/10.51408/issi2025_142

Melting Science: Russian Climate Change Research in the Global Context

Alexey Zheleznov[1], Ekaterina Dyachenko[2], Maxim Dmitriev[3] and Katerina Guba[4]

[1] azheleznov@eu.spb.ru, [2] edyachenko@eu.spb.ru, [3] mdmitriev@eu.spb.ru, [4] kguba@eu.spb.ru
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

Abstract

In this paper we wanted to assess the rapidly growing contribution of Russian scientists to global climate change studies. Our study examines publication patterns and citation impact across national and international journals, based on Scopus data for this area from 2010 to 2023. The analysis highlights shifts in the dissemination of Russian climate-related outputs, reflecting a transition from dominance by national mainstream journals to a more diverse landscape by 2022. Despite Russia's geopolitical isolation and reduced collaboration with the Western academic community, significant contributions to global climate change research persist in recent years. Approximately 90% of citations for Russian-authored articles originate from international journals. Our findings suggest that Russian journals continue to primarily serve the ex-Soviet research community, limiting their broader recognition. The study raises critical questions about the visibility and integration of Russian science in global research agendas. By investigating the interplay between external factors and scientific output, the findings shed light on the evolving role of Russian researchers in addressing pressing global challenges. This scientometric exploration offers insights into how academic isolation influences the structure and impact of national scientific contributions in the context of climate change, with broader implications for the global research ecosystem. A major freezing by western academic community has complicated a national science melting without a potential western diplomatic thaw with Russia as hundreds of regional researchers still prefer national journals.

Introduction

Our research explores Russian climate change science, emphasizing the importance of enabling prospective research to support global sustainability. The climate crisis has catalyzed interdisciplinary scientific efforts worldwide and fostered prominent initiatives promoting international cooperation. In response to the political freezing Russia might reflect on its Soviet-era experience on national self-sufficiency, leveraging existing scientific expertise and human capital (Krasnyak, 2018). Notably, every second Russian author in leading regional journals in Physics and Astronomy has also published at least one article or review in any of the Nature Index journals, that cannot be attributed only to 'low scientific quality' research (Veretennik & Yudkevich, 2023). Despite a Post-Soviet thaw, many Russian scientists continued to prioritize poorly visible local journals, but the rise of the international citation of Russian science has been tied to collaborative publications written directly in English and published in the major international journals (Kirchik et al., 2012). The Western academic community was quite outspoken on breaking academic relationships with Russian universities (Wit & Altbach, 2024). The current wide isolation of Russia's scholars has unfolded an ever-growing crisis that concerns global changes in the Arctic, in particular (Rees & Büntgen, 2024).

The purpose of this research is to examine the international visibility of the Russian climate-related output in both international and national journals from 2010 to 2023. Taking into account the huge variety of known environmental processes on Russia's climate change and the legacy of various academic groups, plenty of Russian researchers are still working in their labs at the time of the ongoing conflict. Notable experts (Oldfield & Poberezhskaya, 2023) warn that the increasing isolation of Russian science from the international community risks deflecting attention away from critical debates in geoengineering and climate modification, thereby alienating this rich scientific tradition at a critical juncture.

As the volume of climate change research grows, scientometric studies have expanded, providing an opportunity to understand a range of issues regarding new knowledge, including the productivity of specific countries and regions such as Central Asia (Vakulchuk et al., 2023). Prior research highlights that papers authored by ecologists from countries where English is a national language attract significantly more citations than those from non-native English speaking countries (Leimu & Koricheva, 2005).

The broad research question quiding this study is: What are the citation patterns of Russian climate change publications, and what do they tell about Russia's integration into global research on this hot topic?

Rather than offering a single assessment, this study aims to present a comprehensive picture that would help us understand whether Russia plays a prominent role in global warming studies. This can help to guess how climate-related sciences would be melted by a significant deterioration in the ties between Russian scientists and the science of Western countries.

The field of climate change research is highly heterogeneous. Russia possesses about 40% of the Arctic region and data obtained by polar scientists are an important element in understanding the rapid changes in global climate. Thus, interdisciplinary cooperation in climate research is a significant element for a deeper and more accurate understanding of climate change. Some works are products of international collaborations, while others are authored by Russian scientists only. To answer the research question we analyzed not just averaged indicators for this diverse flow of publications, but also its constituent parts separately. We believe this approach allowed us to obtain a citation pattern of Russian climatic research.

Material and methods

The metadata for scientific publications authored by Russian researchers were collected from Elsevier's Scopus bibliometric database over a 14-year period, spanning from 2010 at the end of 2023 (last accessed on 27.02.2024). The period was selected due to a significant increase in the volume of publications in this field during these years, and, more important, to the introduction of the so-called "mega-grants" governmental program in 2010, which aimed to establish research laboratories led by prominent scientists. Many representatives of the Russian diaspora, as well as leading domestic scientists, were awarded for their climate change-related projects.

Our search strategy expands upon broader bibliometric approaches, which often rely on standardized keyword-based queries (Fu & Waltman, 2022). However, such general approaches may oversimplify the complexity of climate-related research by overlooking regional terminologies and specific environmental factors relevant to a country's climatic characteristics. To address this limitation, we retrieved scholarly outputs, including papers published in 13 relevant international and national journals as well as those identified using a set of 179 important keywords and expressions[1]. Four types of publications were selected: articles, reviews, letters, and notes. A country restriction was applied — at least one author must have at least one affiliation in Russia.

To better understand relevant details, we categorized research journals into five distinct groups based on their audience orientation:

- International mainstream: the top-100 most influential journals cited by policy documents from the Overton database (Bornmann et al., 2022). These journals serve as key platforms for global scientific discourse and policy-relevant research in general.
- Low quality: journals that were either discontinued from the Scopus database or flagged by 'Beall's List' due to concerns about their editorial practices.
- Russian mainstream: national journals that publish predominantly in English, either as original publications or as translated editions of Russian-language articles.
- Russian non-mainstream: other national journals, often publishing in Russian.
- International non-mainstream: other journals that are neither in the top-tier international category nor classified as low-quality.

Results

Publication Output

Our strategy identified 21556 Russian-authored articles on the topic of climate change. As shown on Figure 1, the annual publication output of Russian authors grew rapidly from 2012. In 2010, Russian authors published 656 papers, and the dataset reveals a remarkable fourfold increase by the end of the period. Figure 1 illustrates the publication's dynamics across different journal categories. Notably, half of these publications are strongly associated with Russian journals.

1 The search query is presented online: https://github.com/OdSt/CLIMATE/blob/master/query.txt

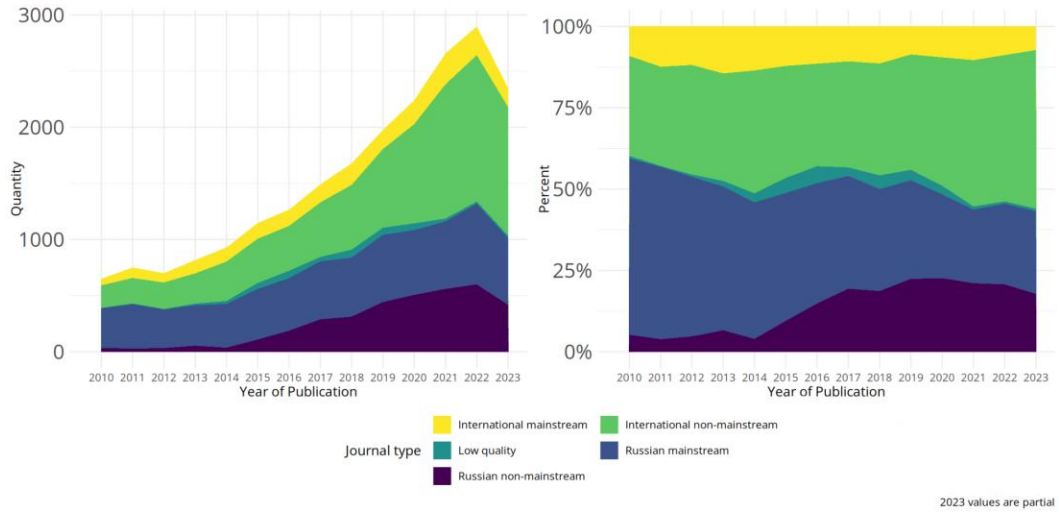


Figure 1. Origin of climate change-related publications by Russian researchers.

Between 2010 and 2014, Russian mainstream journals were the predominant platform for publication, averaging 369 articles annually, which accounted for 48.5% of the total number of articles during this period. After 2014, the share of other types of journals increased (Figure 1b). For instance, while in 2010 Russian non-mainstream journals published 5% of the total number of publications in that year, in 2022 they took up 20%. Another growing trend was demonstrated by the group of international non-mainstream journals. In 2010, this group had 30% of the total number of publications that year, but in 2022, there was already a 45% share. In the group of mainstream international journals, the share of publications remained virtually unchanged and averaged 10%.

Citations

Table 1 shows that different segments of the publication flow of Russian authors are quite different in terms of how many citations they receive and where these recognition come from. We wanted to find out whether the articles in each segment are important to the international community. We put five segments of the publication flow to the rows 2-6 of the Table. Columns from C to G show from which journals citations to Russian articles come from.

We see that the vast majority of citations of Russian articles - about 90% - is received by papers in international journals, not Russian ones. Papers in Russian journals present about half of the publication set, but attract only 11% of citations. Russian papers in international mainstream journals account only for 10% of the publication set, but attract 33% of the citations.

Articles from low quality journals receive not so many citations. Still, when one calculates the citation per paper ratio, it will be higher for low-level journals than for Russian journals (rows 5 and 6). We can confirm previous estimations (Kirchik et al., 2012) that papers in Russian journals (even those published in English) do not attract much attention. Many of these journals publish predominantly the authors from local academic communities.

We have identified that the origin of citations for each category come mostly from journals within the same set. It is interesting to look at the group of papers in Russian journals in English (Russian mainstream journals, line 5). Many of them are translated journals, i.e. they accept manuscripts in Russian and then translate them into English. This significant expenditure of publishing resources serves the goal of making the articles visible to an international audience. Do they achieve this goal according to the citation data? It seems that in general, they do not. The average number of citations of an article in a Russian journal in English is 1.4 over a 3-year window, and this is almost as high as for articles in Russian-language journals - 1.1. For comparison, articles in international journals gain on average 8.1 citations, in high-level international journals - 18.5.

Moreover, 70% of citations to Russian journals originate from Russian journals, indicating that translating articles into English does not significantly enhance their recognition by an international audience.

Several factors may contribute to this citation disadvantage. Lack of journal visibility could be one of them. Russian authors may also be inclined to publish the most interesting results in foreign journals. Another critical factor could be that Russian journals publish few articles created by international teams.

Table 1. Distribution of citations of Russian authors' publications by journal groups.

Papers in	*A) N of Russian papers*	*B) All citations in 3-year window*	*C) Citations from international mainstream journals*	D) Citations from international non-mainstream journals	E) Citations from low-level journals	F) Citations from Russian mainstream journals	G) Citations from Russian non-mainstream journals
All journals	**21556**	**123527**	**25211**	**83098**	**1342**	**9085**	**4656**
International mainstream journals	2197	40602	**14070**	25124	165	830	395
International non-mainstream journals	8385	68376	10510	**53187**	492	2653	1423
Low-level journals	468	980	21	399	**496**	17	46
Russian mainstream journals	6894	9611	531	3321	135	**4871**	749
Russian non-mainstream journals	3612	3958	79	1067	54	714	**2043**

Discussions

This paper has provided a broad overview of the contribution of 25528 Russia-affiliated scientists to global climate change studies. Our findings demonstrate the significant difference in the visibility and impact of papers across five different

groups of journals. Russian scientists, firstly, are actively involved in the study of global perspectives, and secondly, develop a number of locally important issues. Despite long-standing limitations such as insufficient computer capacity, the Russian scientific community has a long history of research on climate modeling and international cooperation in this area (Doose, 2022; Semenov et al., 2024). During freezing times of the Cold War competitions, world-class collaborations on glaciers and sea ice paradoxically melted the Iron Curtain and opened valuable links (Lajus & Sörlin, 2014). We looked at the size of the author teams in each group of the papers, and found that the biggest are those teams that produce internationally co-authored papers in international mainstream journals (median size is 8). Russian-only teams have a median of 3-4 members.

Christine Musselin (2024) describes how world-class researchers strategically navigate between solo and co-authored works, as well as between national and international publication venues. We observe comparable outliers among leading Russian climate scientists. For instance, climate modeler Evgeny M. Volodin has published high-impact solo-authored papers in international mainstream journals (Geophysical Research Letters, 2021; Environmental Research Letters, 2013) while also co-authoring widely cited articles in national mainstream journals (e.g., Izvestiya, Atmospheric and Oceanic Physics, 2010; Russian Journal of Numerical Analysis and Mathematical Modelling, 2018). This dual strategy suggests that top Russian researchers recognize the need to engage both domestic and global audiences, but structural constraints may still limit their international influence.

The broader geopolitical context has increasingly shaped the trajectory of Russian climate change sciences. While current climate mitigation actions are often driven by concerns of economic competitiveness, energy efficiency, and security interests (Kochtcheeva, 2022), the internationally recognized research teams and climate models still serve as vital foundations for climate sciences in Russia. Some scholars argue that Western countries should explore targeted climate policy incentives to sustain engagement with Russian researchers, given the global urgency of climate action (Moe et al., 2023).

If current trends persist, Russian science may face increasing fragmentation, with potential consequences for both national and global climate change research. The context of our study is the deteriorating relations between Russia and most Western countries. This is already affecting and will continue to affect how Russia participates in global climate science.

Acknowledgments

The citations analysis was funded for E.D. and M.D. by a grant from the Russian Science Foundation № 25-28-01490, https://rscf.ru/project/25-28-01490/

References

Bornmann, L., Haunschild, R., Boyack, K., Marx, W., & Minx, J. C. (2022). How relevant is climate change research for climate change policy? An empirical analysis based on Overton data. *PLOS ONE, 17*(9), e0274693. https://doi.org/10/gt3x2q

Doose, K. (2022). Modelling the future: Climate change research in Russia during the late Cold War and beyond, 1970s–2000. *Climatic Change*, *171*(1), 6. https://doi.org/10/gtmtpx

Fu, H.-Z., & Waltman, L. (2022). A large-scale bibliometric analysis of global climate change research between 2001 and 2018. *Climatic Change*, *170*. https://doi.org/10/gss9wz

Kirchik, O., Gingras, Y., & Larivière, V. (2012). Changes in publication languages and citation practices and their effect on the scientific impact of Russian science (1993–2010). *Journal of the American Society for Information Science and Technology*, *63*(7), 1411–1419. https://doi.org/10/f33wnq

Kochtcheeva, L. V. (2022). Foreign Policy, National Interests, and Environmental Positioning: Russia's Post Paris Climate Change Actions, Discourse, and Engagement. *Problems of Post-Communism*, *69*(4–5), 423–435. https://doi.org/10/gqngc7

Krasnyak, O. (2018). *National Styles in Science, Diplomacy, and Science Diplomacy: A Case Study of the United Nations Security Council P5 Countries*. Brill. https://doi.org/10.1163/9789004394445_002

Lajus, J., & Sörlin, S. (2014). Melting the glacial curtain: The politics of Scandinavian–Soviet networks in the geophysical field sciences between two polar years, 1932/33–1957/58. *Journal of Historical Geography*, *44*, 44–59. https://doi.org/10.1016/j.jhg.2013.12.006

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32. https://doi.org/10.1016/j.tree.2004.10.010

Moe, A., Lamazhapov, E., & Anisimov, O. (2023). Russia's expanding adaptation agenda and its limitations. *Climate Policy*, *23*(2), 184–198. https://doi.org/10/gtbw38

Musselin, C. (2024). Transformed Publication Strategies. *European Review*, *32*(S1), S53–S62. https://doi.org/10.1017/S1062798723000546

Oldfield, J. D., & Poberezhskaya, M. (2023). Soviet and Russian perspectives on geoengineering and climate management. *WIREs Climate Change*, *14*(4), e829. https://doi.org/10/gtmtnm

Rees, G., & Büntgen, U. (2024). Russian dilemma for global arctic science. *Ambio*, *53*. https://doi.org/10.1007/s13280-024-02038-z

Semenov, S. M., Mokhov, I. I., Semenov, V. A., Zherebtsov, G. A., & Bardin, M. Y. (2024). Russian science and modern climatology: to the 300th anniversary of the Russian Academy of Sciences. *Fundamental and Applied Climatology*, *10*(1), 5–55. https://doi.org/10.21513/2410-8758-2024-1-05-55

Veretennik, E., & Yudkevich, M. (2023). Inconsistent quality signals: Evidence from the regional journals. *Scientometrics*. https://doi.org/10/gr8kn7

Wit, H. de, & Altbach, P. G. (2024). Navigating University Neutrality in Geopolitical Turmoil: Not So Simple Anymore! *International Higher Education*. https://doi.org/10.6017/895b9e0d.9a48f850

https://doi.org/10.51408/issi2025_137

Missing Links in the *chaîne opératoire* of Citation: The Limitations of Systematic Literature Search in The Social Sciences and Humanities

Kathryn O. Weber-Boer

k.weberboer@digital-science.com
Cornell University, Digital Science (USA)

Abstract

The *chaîne opératoire* is a concept used in archaeology to describe the sequence of interactions with a material object that transform it from raw material to object-in-use, and eventually to disposal. In the context of materiality theory, the concept has been used to identify the varieties of other objects, individuals, and social conditions implicated in the [life] of a seemingly singular object.

The object at the heart of this study is the scholarly literature on the zooarchaeology of the Bronze Age in the South Caucasus. Three corpora are composed, taking three distinct approaches, in order to map the object of study via triangulation. Two alternative methods of systematically exploring the literature are chosen. A genealogical approach (Corpus 1) begins with the "forefathers" of what was then called Transcaucasian archaeology and follows the citations of that work forward. An archaeological approach (Corpus 2) uses a small set of recent publications, highly relevant to the zooarchaeology of the Bronze Age South Caucasus and follows their references backward. The final corpus (Corpus 3) is the bibliography of the first full draft of a doctoral dissertation, assembled by a doctoral student in the US-based tradition of anthropological zooarchaeology (this work is auto-ethnographical). By identifying the citation and collaboration networks within which each corpus is situated, we can reconstruct the temporal, spatial, and interpersonal conditions of knowledge production.

This analysis has two aims. First, to test the application of a systematic approach to the literature of a largely book-based, multi-language field of research. Second, the results should show how well the knowledge represented in the dissertation, as originally drafted, reflected the field, as a whole, and whether there are identifiable research communities with which the research did not initially engage. The three corpora reveal three distinct constellations of actors studying the Bronze Age in the South Caucasus. The context of archaeological knowledge production includes the history of archaeological practice in Europe and Southwest Asia, relationships of status, and resource inequality. A systematic approach may expose the implicit geographic, temporal, and institutional patterns of knowledge production influencing the dissertation, highlighting the traditions the dissertation draws from, the discourses it contributes to, and the literature missing from consideration. A comparison of these corpora reveals the consequences of these conditions, identifying scholars and traditions of scholarship, to chart a landscape of scholarship that extends beyond what is "known" according to standard practices in archaeological research. This comparison, thus far, simultaneously highlights the limitations and the advantages of systematic literature review and the usual, ad hoc approach.

This work consists of both qualitative and quantitative analysis, which is called for by the irregular nature of the underlying data. The quantitative analysis consists of an overlap analysis, of the publications most cited within each corpus. The qualitative assessment of the results draws upon the history of archaeological practice, research funding, and social networks within which these publications and citations are situated.

Introduction

This research in progress considers qualitative factors and systematic, quantitative analysis to approach the background literature for anthropological archaeological

research into the effect of human-animal relationships on sociopolitical organization and the establishment of political authority in the South Caucasus during the Bronze Age. I combine systematic approaches to literature selection with auto-ethnography, reporting the process by which a doctoral student in the US-based tradition of anthropological zooarchaeology assembles a bibliography. This simultaneously highlights the limitations of bibliometric approaches to systematic literature review in a book-based, multi-language field, and the limitations of the usual, *ad hoc* approach.

To approach the universe of scholarship relating to the subject of the dissertation, I compile three corpora, from which I derive three overlapping constellations of actors. Later work will draw on information about these scholars, and the concepts that their research has covered. This research asks how well the knowledge represented in each of these approaches reflect the field as a whole, and which scholarly communities are lost in each.

Archaeological knowledge production takes place in the context of the history of archaeological practice in Europe and Southwest Asia, relationships of status between senior and junior scholars, and geographically determined resource inequality (including both research funding and time). A systematic approach is well-suited to recording the consequences of these conditions (where that scholarship which engages most actively with an international community becomes more visible), and it also has the potential to chart a landscape of scholarship that extends the boundaries of the "known world". It explores the geographic, temporal, and institutional patterns of knowledge production influencing archaeological research, to articulate clearly what traditions the dissertation as originally drafted had drawn from; what discourses it contributed to; and what literature was missing. What such a systematic approach may miss is scholarly literature produced with different practices of formatting, publication, and dissemination.

Background

Approaching this study requires understanding the practices of research output in the social sciences and humanities, the background of the digital resources available, and the history of archaeological knowledge production in the region of the South Caucasus.

The social sciences and humanities are a challenge to digital bibliometric datasets. They often rely on non-article research outputs (e.g., books and conference presentations), not all of which are indexed, and multiple languages continue to be used for scholarly communication. In archaeological research local/regional archaeological publications are essential to knowledge dissemination and archives may contain one of few, or the only, example of research materials.

The Dimensions dataset includes one of the largest collections of metadata about research published around the world, with no geographic or temporal exclusions (the earliest publications date to 1665), as long as the metadata are digitized and made available either openly or through data sharing agreements. In principle, all languages are included. That said, there are limitations for all scholarly metadata providers. The most relevant limitation for the study presented here is a paucity of

scholarly output published in Russia. Despite the existence of Russian-language online digital libraries and metadata repositories, access to these resources has been blocked by the current geopolitical landscape, including governmental sanctions prohibiting contracts with Russian entities, and ethical constraints emerging from Russian national policy, including the situation in Ukraine. For example, eLibrary.ru contains over 70 million articles and 1.7 million books, and—according to its search interface, conference materials, dissertations, grants, and datasets, but it cannot be linked to existing databases. In the past, access was complicated by uneven adoption of CC0 licensing and open science expectations, which impacted both business models and infrastructure development. Clarivate had absorbed these data in the past, to sell the Russian Science Citation Index (PR Newswire, 2014), but that dataset is now defunct for geopolitical reasons (Scientific Publications, 2024). In considering other data sources, Lens.org had a smaller quantity of Russian-affiliated publications. OpenAlex and Dimensions have partially overlapping Russian-affiliated publications, with over 1.8 million publications in common. [Identify the fields where Dimensions has non-OA sources, and where OA has non-Dimensions sources?]

The origins of archaeology in the South Caucasus, as we would recognize it today, coincided with the completion of Russian imperial control of the region. This established a familiar relationship between political, social, and scientific knowledge which would persist well into the 21st century. The earliest published work, in the 1880s, was produced by Jacques de Morgan. An excellent educational system and access to reliable and sufficient resources produced generations of archaeologists, who were able to set, challenge, and test chronologies; explain technological and social innovations; and archaeologists in the region recorded remains from every period of human occupation, including the oldest hominid outside of Africa. Twice. The collapse of economic and political order in the 1990s led to twenty years during which young archaeologists found it difficult to find a professional position after their disciplinary training. Because stability returned without significant economic improvement, archaeologists in the South Caucasus became heavily dependent on research funded by international grants, from the US, Australia, Germany, France, and Italy. This led to shifts in the research agenda, and reevaluation of long-established facts (from chronological frameworks to the very idea of socioeconomic progress). My own work is embedded in that period, having begun in the early 2010s, when I could count on one hand the number of professional junior archaeologists in Armenia and Georgia. My academic and disciplinary training in the US drove me to identify explanations and logics that were unconvincing to my understanding of narratives of sociopolitical change. This gap-finding, historically critical approach to knowledge production risks the alienation of colleagues raised in different epistemological traditions and the loss of the intellectual labor of our scholarly predecessors.

Fortunately, several successful long-term partnerships of European, American, and Australian archaeologists with established figures in Armenian and Georgian archaeology have entailed respectful knowledge exchange. Although the attrition rate of young archaeologists between their studies and professional employment

remains high (as it does for students of archaeology globally), it seems to be declining. An increasing number of junior archaeologists engage in fieldwork and publication, are granted responsibility as primary investigators on government permits, and--most importantly—are more frequently employed, for example by the Georgia National Museum. It is worth noting explicitly the funnel of this narrative of knowledge production: moving from the South Caucasus, to the countries of Armenia and Georgia, to the employment practices of a single institution. This illustrates another relevant condition of research in the area, which in experience can often be practically limited to a very small region.

As for zooarchaeology, the party traditionally responsible for recording and analyzing faunal material in the South Caucasus has been the paleozoologist. This has had the effect of reliable registration of the presence and absence of species occurring in older archaeological sites, but a lack of the kind of detail that an anthropological zooarchaeologist generally relies upon. The discipline of zooarchaeology has grown, and the importation by international teams of specialists from Australia, France, Germany, Italy, and the US has meant a growing trove of faunal material and increasingly detailed records. However, a diversity of professional training has meant that many of the resulting datasets are to some degree incompatible. Further, a tendency for zooarchaeologists to be found among graduate students rather than funded as highly skilled specialists, has meant that participation is often fleeting.

Methods

Two alternative methods of exploring the literature are systematic. One takes a genealogical approach, beginning with the "forebearers" of what was then called Transcaucasian archaeology, with a focus on faunal remains from the Bronze Age (method adapted from Garfield 2002). This approach follows these scholars forward through the literature, seeking the researchers by whom, and alongside whom, they are referenced. Two approaches are taken to find these references: primary references were found via a full-text search for in-text citation and secondary citations used these primary references to search for identifiers in the metadata. Corpus 1 is the resulting publication list (the candidate population).

The second approach could be called archaeological. A search of the literature for papers similar to the dissertation produces the most visible surface of zooarchaeology of the Bronze Age South Caucasus. These publications are the latest structures built on a mound of thought, research, and labor. Corpus 2 was constructed by following their references backwards, in two steps (mirroring the steps forward of Corpus 1). The document set assembled by this approach is Corpus 2.

Finally, the third corpus is composed of the bibliography which was submitted with the 2019 draft of the author's doctoral dissertation (in progress), *The Herd, the Hearth, and the Hunt: Human-Animal Relationships in the Bronze Age South Caucasus*. The discovery period of this dissertation could not be properly called systematic. Disciplinary training in anthropological archaeology is shaped by the constellation of mentors, courses of undergraduate and graduate study, and

professional networks that a junior scholar builds in the field (and by serendipity). This can be considered a *discipline-network* approach.

From corpora 1 and 2, the unique authors of publications with more than one citation within their corpus are extracted. These authors are then situated in the citation and collaboration networks, to reconstruct temporal, spatial, and interpersonal conditions of knowledge production. By putting these two approaches in conversation, it should be possible to discover the places the standard approach did not lead and to identify divergent tendencies in the field as it appears today.

The results of these three approaches are then compared. First the individual researchers are extracted from each corpus. This step is taken to reduce the impact of inconsistent coverage of output in multiple languages.

Results

Corpus 1

Some of the founding scholars of the study of the Bronze Age in the South Caucasus (putting a somewhat artificial upper boundary of 1950), are Jacques de Morgan (1889), Nikolai Marr (1894, 1922), Iessen (1935), Boris Kuftin (1941, 1944, 1949), and Piotrovskii (1944, 1949). For zooarchaeology in Georgia and Armenia, Oleg Bendukidze and Nina H. Manaserian are the major early figures, respectively.

A full-text search for the combination of each author's name and the year of publication given above, within two-word proximity, in the full text of all publications in the Dimensions publication dataset published before 2020 resulted in 309 unique publications. There were no publications found for the N. Manaserian who was the scholar responsible for early Armenian zooarchaeology, though there were several papers found by her daughter, N. Manaserian (or Manaseryan), which were added to Corpus 2. By adding the publications which have cited those 309 publications, Corpus 1 is composed of in a total of 1380 unique publications.

Corpus 2

A search in the titles and abstracts of the Dimensions publication dataset for ("South Caucasus" [within two-word proximity]) OR ("Kura Araxes") OR ((Caucas*) AND (Armenia OR Georgia OR Azerbaijan)) OR (Transcauc*)) AND ("Bronze Age") AND (zoomorph* OR (((archaeozoolog* OR zooarchaeolog*) OR (arch* AND (fauna* OR animal*))))) produced 22 results from the Dimensions API, each of which was published from 2015-2025. The publications authored by N. Manaserian and N. Manaseryan in the Dimensions dataset were added to these, as were the 500 documents produced by a similarity search using the abstract of the dissertation. This resulted in a total of 531 publications, and when the references listed in each of those publications were added to the corpus, Corpus 2 is composed of 3769 unique publications.

Corpus 3

The third corpus consists of the items of the bibliography produced through what I have termed the discipline-network approach. The process of compiling this

literature began with the composition of a reading list for what Cornell University terms the "B Exam", which is a precursor to the dissertation proposal, and completion of which confers a Master's degree. The reading list began with familiar texts assigned in introductory coursework (introductions to archaeology theory and methods, the region of the South Caucasus, landscape and mortuary archaeology, and social zooarchaeology). Additions to the list were identified using the references sections of these texts, the suggestions of the doctoral committee (Adam T. Smith, Lori Khatchadourian, and Nerissa Russell), and through conversation with local archaeologists in Armenia and Georgia during fieldwork performed in these countries. There were 170 referenced publications by 146 unique authors (only including the first author of publications with 3 or more authors).

The last names of researchers who authored the publications in Corpus 3 joined with the authors of the subset of the Corpus 1 and Corpus 2 published before 2020, which had more than two references from either corpus 1 or 2, produced a list of 1227 distinct researchers. The overlap of these researchers by corpus can be seen in Figure 1.

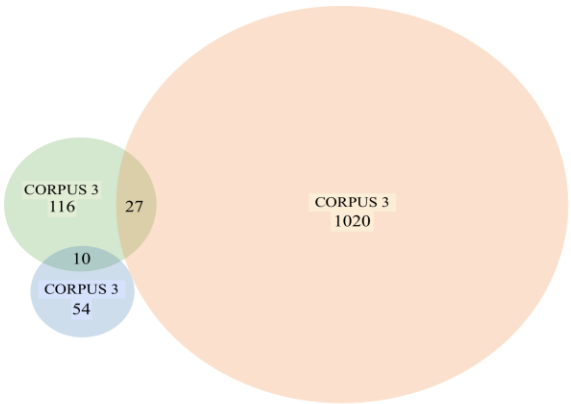


Figure 1. Unique authors per corpus.

Discussion

It is too easy for scientometricians to sidestep the limitations of our datasets and approaches to the humanities and (some of the) social sciences. Focusing on the fields which are best suited to systematic analyses because of the completeness of coverage and the homogeneity and regulation of research design and output can lead to an inaccurate sense of the reliability, validity, and superiority of such approaches. The importance of systematic review as a collection of all relevant literature on a subject is well established, but it is equally true for archaeological studies that a missing reference makes the difference between competent and insufficient research. It becomes essential, then, to consider how to address a subject of research for which available datasets are incomplete, whether due to missing languages, "non-traditional" research outputs, or divergent practices of digitization.

Conclusions

The research thus far shows a surprising variation in the results of each approach. The isolated nature of the second corpus suggests that either Corpus 1 and 2 require a tertiary level of citation/reference (that is, the references to the secondary citations of Corpus 1 and the citations of the secondary references of Corpus 2), or that some additional constraint by subject should be applied to Corpus 2. After refining the approach, the next step of research (inspired by Leydesdorff, 2010) will involve looking at trends in concepts over time, the geographic distribution of citing and cited authors, and trends in the academic age of the authors citing and cited in each corpus.

Acknowledgments

The author would like to acknowledge the support of Digital Science, which provided access to the Dimensions dataset upon which this work depended.

References

Bendukidze, O. (1979). Holocene Vertebrate Fauna of Georgia, Tbilisi, 1-115.

Garfield, E., Pudovkin, A.I., & Istomin, V.S. (2002). Algorithmic Citation-Linked Historiography—Mapping the Literature of Science. ASIST, 14-24.

Iessen, A.A. (1935). Iz istorii drevneishei metallurgii Kavkaza [Of the history of ancient metal work in the Caucasus]. GAIMK.

Kuftin, B.A. (1949). Arxheologia Kavkasia [Archaeology of the Caucasus].

Kuftin, B.A. (1941). Raskopki v Trialeti [Excavations in Trialeti]. Tbilisi.

Leydesdorff, L. (2010). Eugene Garfield and Algorithmic Historiography: Co-Words, Co-Authors, and Journal Names. *Annals of Library and Information Studies*. …

Manaserian, N. (1986). Spreading and economy utilization of wild and domestic representatives Ovis and Capra family. *Zoological Papers*, 20:80-99.

Marr, N.J. & Orbeli, I.A. (1922). Arxeologicheskaya ekspeditsia 1917 [Archaeological expeditions of 1917]. CPB.

Morgan, J. de (1889). *Mission scientifique au Caucase études archéologiques & historiques.* https://doi.org/10.24157/arc_12444

Morgan, J. de (1889). Anneaux-monnaies du Caucase et de l'Armenie. *Comptes-rendus des séances de l'année – Académie des inscriptions et belles-lettres*, 33(4), 263-264. https://doi.org/10.3406/crai.1889.69673

Piotrovskii, B.B. (1949). Arxeologia Zakavkazia. [Archaeology of Transcaucasia].

Piotrovskii, B.B. (1944). Istoria i kultura Urartu. [The History and Culture of Urartu]. Yerevan.

PR Newswire. (2014). *Thompson Reuters Collaborates with Russia's Scientific Electronic Library eLibrary.RU to Showcase Nation's Leading Research in Web of Science*. 1 October. https://www.prnewswire.com/news-releases/thomson-reuters-collaborates-with-russias-scientific-electronic-library-elibraryru-to-showcase-nations-leading-research-in-web-of-science-277730241.html. Accessed 31 January 2025.

Scientific Publications. (2024). Russian Science Citation Index excluded from Web of Science: What does it mean for Kazakhstan? https://spubl.kz/en/blog/rossysky-indeks-nauchnogo-tsitirovaniya-isklyuchen-iz-web-of-science-chto-eto-znachit-dlya-kazakhstana. Accessed 31 January 2025.

https://doi.org/10.51408/issi2025_145

National Mobility and Career Performance of the Scientific Workforce in Colombia

Jesús María Godoy[1], Yajie Wang[2], Julián D. Cortés[3]

[1]*jesus.godoy@unibague.edu.co*
International Bussiness Administration, Universidad de Ibagué, Cra. 22 – Cll. 67, Ibagué (Colombia)

[2]*yajie.wang@uni-corvinus.hu*
Center for Collective Learning, Corvinus Institute for Advanced Studies (CIAS), Corvinus University, 1093 Budapest (Hungary)

[3]*julian.cortess@urosario.edu.co*
School of Management and Business, Universidad del Rosario, Autopista Norte Cll. 200, Bogotá (Colombia)

Abstract

Academic mobility plays a crucial role in fostering intellectual collaboration, knowledge transfer, and the internationalization of science. While existing research has extensively examined international scientific migration, national mobility, particularly in middle and low-income countries, remains underexplored. This study investigates the relationship between national mobility and career performance within Colombia's scientific workforce, using unique data from national assessments conducted between 2013 and 2021. The analysis includes 12,084 researchers, with career trajectories evaluated using ordered probit regression models. Here, we show that mobility, defined as moving across municipalities, is not significantly associated with changes in researchers' career rankings. Instead, structural and individual factors, such as prior rank, institutional affiliations, and years of experience, emerge as the main drivers of career progression. Additionally, residing in large cities appears to negatively affect rankings, possibly due to intensified competition for resources and funding. These findings highlight the importance of cumulative advantage mechanisms and institutional dynamics over geographic mobility in shaping scientific careers. Future research should expand on this framework by incorporating institutional prestige and cross-country comparisons to better understand the nuanced interplay between mobility and academic performance.

Introduction

Academic mobility is a key factor in science policy, promoting intellectual collaboration, innovative knowledge production and transfer, and the internationalization of national science systems (Cavalli & Teichler, 2015; Gureyev et al., 2020; Momeni et al., 2022; Morano-Foadi, 2005; Soete et al., 2021; Sugimoto et al., 2017). It has also become a prominent subject in quantitative science studies, where the most relevant topics include the development of methodological approaches, the flows of scientific migration, the impact of scientific mobility, factors driving scientific mobility, and historical perspectives (Gureyev et al., 2020). Despite this, most research has been directed toward international migration, highlighting the lack of understanding of the dynamics of national scientific workforce mobility in middle and low-income countries (Liu et al., 2024).

Migration and mobility, though related, differ primarily in their permanence. According to Teichler (2015), migration signifies a permanent relocation, such as a scientist moving from the country of citizenship to another to take up a permanent research position. In contrasts, mobility refers to non-permanent or repeated movements without a permanent change in residence, such as a scientist participating in an international research exchange for a few months (Teichler, 2015).

Concerning exceptional studies on national mobility, evidence from the United States shows that professors are more likely to move to institutions with higher research intensity and from rural to urban areas, with female professors more frequently relocating within the same geographic region than their male counterparts (Yan et al., 2020). At Washington State University, researchers showed high mobility rates, with domestic movers demonstrating greater citation impacts than international movers (Payumo et al., 2018). In Italy, the centralized and non-competitive university system results in lower post-mobility performance, particularly for less productive researchers (Abramo et al., 2022). Faculty in Turkish public universities, especially women, older individuals, and those in major cities or well-established institutions, are less likely to move nationally (Yuret, 2023). Tuning the attention into Latin America, the signing of NAFTA increased the flow of inventors in México to multinational companies, exacerbating the brain drain, while regional disparities in mobility persist, with Mexico City as a key destination (Aboites & Díaz, 2018). Over time, migration intensity has decreased, but the diversity and density of migration networks across Mexican states has increased (Miranda-González et al., 2020).

In this context, this research aims to test the following hypothesis for the Colombian scientific workforce:

H_0: Researchers' mobility is not associated with career performance (i.e., changes as progressions or declines in their national assessment ranking).

H_a: Researchers' mobility is associated with career performance.

We choose Colombia as a Latin American country with notable characteristics such as a science system with negligible financial resources but noticeable efficiency in scientific output and a history of intense forced internal displacement caused by an armed conflict since the 1960s (Cortés & Ramírez Cajiao, 2024; SCImago, 2020; UNHCR, 2023). In pursuing this goal, we aim to contribute to the emerging literature on national academic mobility by examining the relationship between mobility and academic career dynamics, leveraging open-access data from government agencies rather than relying solely on traditional bibliographic sources like WoS or Scopus.

Methodology

Data

We used open-access datasets curated and issued by the Colombia's Ministry of Science, Technology, and Innovation (MinCiencias). These datasets provide information from national assessments conducted in 2013, 2014, 2015, 2017, 2019, and 2021. Our analysis sources the socioeconomic data and academic career data on Colombian researchers (MinCiencias, 2023).

Colombian researchers are assessed as units within the national evaluation system, categorized into *Junior, Associate, Senior,* or *Emeritus* ranks based on criteria like academic output, leadership, and mentoring. Researchers are responsible for updating their portfolios on the national platform (CvLAC), detailing their disciplinary expertise and research outputs, with oversight from institutional and research group leadership. However, rank progression does not influence salary or career advancement within their employing institutions (Vasen et al., 2023).

We sub-sampled the data to the cohorts of researchers assessed in 2013, 2014, and 2015 which enables at least a 5-year window —a standard time window in research evaluations (Wang, 2013)— to examine the dynamics of mobility and career advancement. We excluded researchers who migrated to another country at least once during any national assessment, as our focus is solely on domestic mobility. However, researchers born abroad but currently residing in Colombia were included in the sample. This sub-sample comprises 12,084 researchers.

Methods and variables

In this exploratory stage of the project, we implemented an *ordered probit regression* to analyze the association between researcher mobility patterns and their career performance. The dependent variable is the last ranking observed for each researcher. Besides mobility-related variables, we also included additional demographic, institutional, and geographic variables into the model as independent and control variables.

Demographic variables.

- Gender: a proxy to assess potential disparities or biases in rankings and career progression based on gender.
- Age: a proxy for capturing career stage, maturity, or productivity levels, which can influence ranking changes.

Mobility and geographic factors.

- Mobile researcher (*moved*): the municipality of residence differs from the municipality of birth (1 = yes, 0 = no).
- Number of cities (*n_cities*): number of different cities of residency of the researchers between 2013-2021. A proxy for a researcher's frequent mobility patterns/intensity (i.e., adaptability or flexibility) (1 city=~95% of researchers; 2 cities=~4%; 3 cities=~1%)
- Living in a big city *(pop_gt1M):* Dummy variable indicating whether the researcher resided in a city with more than 1 million inhabitants in 2020 (1 = yes, 0 = no). Large cities typically offer better research infrastructure, networking opportunities, and resources.

Ranking and institutional characteristics.

- Ranking first (*ranking_first*): the first rank at which the researcher was assigned in the national assessment.

- Ranking changes: number of changes in the researchers rank in the sample. Zero (0) would be those who never changed rank. A proxy for a researcher's (in)stability in their career progression or decline.
- Number of institutional affiliations (*institution_id_n*): corresponds to the number of different institutions to which the researcher is attached. A proxy for tracking a researcher's changes in research environments and diversity of potential academic collaboration.

Academic career.

- Career performance *(progress_career): c*orresponds to the number of changes —positive: progression(s), negative: decline(s)— in the researcher's career. A proxy for measuring the career trajectory/dynamic based on changes in their national rank.
- Career upward *(progress_carreer_dummy): d*ummy variable that corresponds only to progressions in the researcher's career, useful for identifying drivers of career advancement and enabling a separate evaluation of progression versus stagnation or decline.
- Experience: defined as the years since the researcher first participated in the calls (2021 - [year of the first call]). A proxy that reflects cumulative experience and academic visibility.

Supplementary material 1 reports the descriptive statistics of the variables.

Results

The Table 1 reports the results, of which we will focus on the *Ranking Last III* model which shows the highest explanatory power and used a wider range of independent variables. This model's McFadden's pseudo R-squared indicates that ~10% of the variation in researchers' career rankings is explained by the included variables. The results largely support the null hypothesis (H_0), which posits that researchers' mobility is not associated with career performance. The variable `moved` shows no significant relationship across all models, indicating that moving institutions does not predict changes in ranking. Similarly, the significance of `n_cities` in the first model diminishes when additional controls are included, suggesting its limited explanatory power. Instead, the analysis highlights the importance of structural and individual characteristics—such as `ranking_first`, `experience`, and `institution_id_n`—as primary drivers of career performance. Additionally, contextual effects, captured by regional variables like `pop_gt1M`, play a significant role, further diminishing the role of mobility in explaining changes in researchers' rankings.

Table 1. Ordered probit regression results.

Variable	Ranking Last I	Ranking Last II	Ranking Last III
gender	0.1494***	0.1457***	0.1442***
	(0.0222)	(0.0222)	(0.0222)
age	-0.0040***	-0.0025**	-0.0022**

		(0.0011)	(0.0011)	(0.0011)
n_cities		0.0988**	0.0096	0.0070
		(0.0469)	(0.0478)	(0.0478)
ranking_first		0.8483***	0.8567***	0.8598***
		(0.0188)	(0.0188)	(0.0188)
moved		0.0023	-0.0041	-0.0164
		(0.0218)	(0.0218)	(0.0221)
experience		0.2332***	0.2277***	0.2291***
		(0.0147)	(0.0147)	(0.0147)
	1/2	2.8554***	3.0297***	3.0084***
		(0.1232)	(0.1246)	(0.1247)
	2/3	-0.2095***	-0.2038***	-0.2028***
		(0.0160)	(0.0160)	(0.0160)
	3/4	0.4433***	0.4480***	0.4485***
		(0.0192)	(0.0192)	(0.0192)
institution_id_n			0.1703***	0.1731***
			(0.0166)	(0.0166)
pop_gt1M				-0.0821***
				(0.0216)
Log-likelihood ratio chi-squared		2662.16	2767.18	2781.69
Log-likelihood ratio p-value		0.0000	0.0000	0.0000
McFadden pseudo R squared		0.0992	0.1032	0.1037
Obs		12084	12084	12084

Note: p<.1, ** p<.05, ***p<.01

Discussion and conclusions

This study aimed to contribute to the emerging literature on national academic mobility by examining the relationship between mobility and academic career dynamics. The results align with the null hypothesis (H_0), suggesting no significant association between researchers' mobility and career performance. Frequent changes in residency, even at the national level, might create a perception of geographic instability, which could influence upward career performance efforts. Instead, the cumulative advantage and path dependency—captured by variables such as ranking_first and institution_id_n—provide a stronger explanation for the latest career rankings, particularly among seasoned researchers with extensive experience (Merton, 1988; Price, 1976). Furthermore, the inclusion of contextual variables, such as pop_gt1M and its negative effect, likely reflects the heightened competition for funding and talent in large municipalities and cities, which often serve as key hubs for attracting researchers (Verginer & Riccaboni, 2021). Our study is limited to a single national case and does not account for further institutional factors (e.g., directional mobility towards reputable national institutions), or regional/national variables (e.g., socio-economic factors acting as push/pull drivers for mobility). Future stages of the project will incorporate some of these variables and expand the

analysis to include comparative cases, potentially, from other developed/developing countries. Also, it will incorporate additional variables and expand the analysis to include comparative cases from other developed and developing countries while also exploring alternative indicators of mobility, such as institutional changes within the same city or research collaborations across institutions to capture a more nuanced understanding of mobility's impact on career performance, and disentangling the mechanisms underlying the negative impact of large cities, including competition, resource distribution, and policy-making priorities.

Acknowledgments

The authors express their gratitude to Giovanni Abramo and Andrea D'Angelo for their insightful comments and constructive feedback, which have significantly enriched this paper.

Supplementary material

Supplementary material 1 Descriptive statistics

	Total	Moved	Not Moved
N	12,084	6,604 (54.7%)	5,480 (45.3%)
Variable	Total	Moved	Not Moved
gender (= male)	7,668 (63.5%)	4,314 (65.3%)	3,354 (61.2%) ***
Age	46.073 (10.306)	46.310 (10.587)	45.787 (9.949) *
experience	7.392 (0.772)	7.395 (0.775)	7.388 (0.768)
n_cities	1.052 (0.232)	1.095 (0.307)	1.000 (0.000) ***
ranking_changes	0.791 (0.919)	0.793 (0.927)	0.787 (0.909)
ranking_first	1.334 (0.582)	1.330 (0.582)	1.339 (0.582)
ranking_last	1.756 (0.868)	1.758 (0.872)	1.753 (0.864)
institution_id_n	1.354 (0.643)	1.390 (0.680)	1.309 (0.593) ***
moved (= 1)	6,604 (54.7%)	6,604 (100.0%)	0 (0.0%) ***

') Kruskal-Wallis test, ") Fisher exact test, "") Chi-Square test. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Standard deviation in parenthesis

References

Aboites, J., & Díaz, C. (2018). Inventors' mobility in Mexico in the context of globalization. *Scientometrics*, *115*(3), 1443–1461. https://doi.org/10.1007/s11192-018-2645-6

Abramo, G., D'Angelo, C. A., & Costa, F. D. (2022). The effect of academic mobility on research performance: The case of Italy. *Quantitative Science Studies*, *3*(2), 345–362. https://doi.org/10.1162/qss_a_00192

Cavalli, A., & Teichler, U. (2015). Preface - Issue on Mobility and Migration in Science. *European Review*, *23*(S1), 1–5. https://doi.org/10.1017/S1062798714000751

Cortés, J. D., & Ramírez Cajiao, M. C. (2024). The policy is dead, long live the policy—Revealing science, technology, and innovation policy priorities and government transitions via network analysis. *Quantitative Science Studies*, *5*(2), 317–331. https://doi.org/10.1162/qss_a_00295

Gureyev, V. N., Mazov, N. A., Kosyakov, D. V., & Guskov, A. E. (2020). Review and analysis of publications on scientific mobility: assessment of influence, motivation, and trends. *Scientometrics*, *124*(2), 1599–1630. https://doi.org/10.1007/s11192-020-03515-4

Liu, T., Pei, T., Fang, Z., Wu, M., Liu, X., Yan, X., Song, C., Jiang, J., Jiang, L., & Chen, J. (2024). Spatiotemporal mobility network of global scientists, 1970–2020. *International Journal of Geographical Information Science*, *38*(10), 1991–2018. https://doi.org/10.1080/13658816.2024.2369540

Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, *79*(4), 606–623. https://doi.org/10.1086/354848

MinCiencias. (2023). *La Ciencia en Cifras*. https://minciencias.gov.co/la-ciencia-en-cifras

Miranda-González, A., Aref, S., Theile, T., & Zagheni, E. (2020). Scholarly migration within Mexico: analyzing internal migration among researchers using Scopus longitudinal bibliometric data. *EPJ Data Science 2020 9:1*, *9*(1), 1–26. https://doi.org/10.1140/EPJDS/S13688-020-00252-9

Momeni, F., Karimi, F., Mayr, P., Peters, I., & Dietze, S. (2022). The many facets of academic mobility and its impact on scholars' career. *Journal of Informetrics*, *16*(2). https://doi.org/10.1016/j.joi.2022.101280

Morano-Foadi, S. (2005). Scientific Mobility, Career Progression, and Excellence in the European Research Area1. *International Migration*, *43*(5), 133–162. https://doi.org/10.1111/j.1468-2435.2005.00344.x

Payumo, J. G., Lan, G., & Arasu, P. (2018). Researcher mobility at a US research-intensive university: Implications for research and internationalization strategies. *Research Evaluation*, *27*(1), 28–35. https://doi.org/10.1093/reseval/rvx038

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/https://doi.org/10.1002/asi.4630270505

SCImago. (2020). *SJR - International Science Ranking*. https://bit.ly/3Ih0qMa

Soete, L., Schwaag, S., Stierna, J., & Hollanders, H. (2021). European Union. In UNESCO Science Report 2021 (Ed.), *UNESCO Science Report 2021* (pp. 254–289). United Nations. https://doi.org/10.18356/9789210058575

Sugimoto, C. R., Robinson-Garcia, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature*, *550*(7674), 29–31. https://doi.org/10.1038/550029a

Teichler, U. (2015). Academic Mobility and Migration: What We Know and What We Do Not Know. *European Review*, *23*(S1), S6–S37. https://doi.org/10.1017/S1062798714000787

UNHCR. (2023). *Colombia situation*. https://reporting.unhcr.org/operational/situations/colombia-situation

Vasen, F., Sarthou, N. F., Romano, S. A., Gutiérrez, B. D., & Pintos, M. (2023). Turning academics into researchers: The development of National Researcher Categorization Systems in Latin America. *Research Evaluation*. https://doi.org/10.1093/RESEVAL/RVAD021

Verginer, L., & Riccaboni, M. (2021). Talent goes to global cities: The world network of scientists' mobility. *Research Policy, 50*(1). https://doi.org/10.1016/j.respol.2020.104127

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9

Yan, E., Zhu, Y., & He, J. (2020). Analyzing academic mobility of u.S. professors based on orcid data and the Carnegie classification. *Quantitative Science Studies, 1*(4), 1451–1467. https://doi.org/10.1162/qss_a_00088

Yuret, T. (2023). Predicting mobility and research performance of the faculty members in the economics departments at Turkish public universities. *Quantitative Science Studies, 4*(1), 167–185. https://doi.org/10.1162/qss_a_00238

https://doi.org/10.51408/issi2025_128

National research, national policy: how local research fuels Brazil's policy

Bernardo Cabral[1], Evandro Cristofoletti[2], Karen Esteves Fernandes Pinto[3], Sergio Salles-Filho[4], Yohanna Juk[5]

[1] *bernardopcabral@gmail.com*
Federal University of Bahia (Brazil)

[2] *evcoggo@unicamp.br*, [3] *karenefp@unicamp.br*, [4] *sallesfi@unicamp.br* ,
[5] *yohannajuk91@gmail.com*
State University of Campinas (Brazil)

Abstract

Brazil's policymaking has long relied on scientific evidence to address its socio-economic, environmental, and public health challenges. However, persistent budget cuts and political challenges, particularly during recent administrations, have severely impacted the nation's scientific research ecosystem. Despite these setbacks, the Brazilian government continues to integrate both domestic and international research into its policies, highlighting the resilience and relevance of its scientific community. This study investigates the extent to which scientific research, particularly domestic outputs, informs Brazilian government policy. By analyzing policy documents and their citations, we aim to understand the role of local and international research, key funding agencies, and research institutions in shaping Brazil's public policies. We utilized the Overton database, one of the largest repositories of global policy documents, to analyze over 100,000 policy documents published by Brazilian government institutions. These documents were cross-referenced with the Web of Science database to identify 35,000 cited research papers. Citations were categorized by language, geographical origin, funding agency, and institutional affiliation. The data were then evaluated to identify patterns and trends in the use of research by Brazilian policymakers. The analysis reveals that 95.5% of cited research in Brazilian policy documents is published in English, with only 4.1% in Portuguese. International research, particularly from the United States, dominates, accounting for 13,994 articles, while Brazilian research ranks second with 6,350 articles. Domestic institutions, such as the University of São Paulo (USP), State University of Campinas (UNICAMP), and the Federal University of Rio de Janeiro (UFRJ), feature prominently in policy citations, demonstrating their critical role in producing locally relevant research. The study also highlights the importance of funding agencies, with domestic institutions like the National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), and the São Paulo Research Foundation (FAPESP) leading in support of cited research. Despite these contributions, international agencies, including the National Institutes of Health (NIH) and the National Science Foundation (NSF), play a significant role in funding research that informs Brazilian policy.

Introduction

The Brazilian research ecosystem has experienced severe financial constraints since 2014, marked by continuous cuts in federal funding that have significantly hindered the country's capacity for scientific production. By 2017, the budget for the Ministry of Science, Technology, and Innovation (MCTIC) had been slashed by 44%, reaching its lowest level in over a decade, a trend consistent with broader austerity

measures that crippled federal and state-level research funding (Angelo, 2016, 2017; Gibney, 2015). These budgetary cuts have severely impacted ongoing research and technological projects, leaving institutions struggling to maintain operations (Moutinho, 2022).

Despite these financial challenges, science continues to play a crucial role in shaping Brazilian policies. Institutions like the Oswaldo Cruz Foundation (Fiocruz) have provided key research to guide public health responses, such as during the Zika virus and COVID-19 crises. Similarly, the Ministry of the Environment relies on research from the National Institute for Space Research (INPE) to address deforestation and environmental preservation. Meanwhile, state-level initiatives, particularly in São Paulo, have continued to drive scientific innovation, leveraging the region's unique capacity for applied and multidisciplinary research (Faleiros, 2018). The same can be said for the role of the Brazilian Agricultural Research Corporation (Embrapa) in agricultural research.

However, the broader context of science denialism, compounded by politically driven narratives under past administrations, has further undermined the credibility of scientific expertise in policymaking (Diele-Viegas et al., 2021). The current study explores how scientific evidence, both domestic and international, is integrated into Brazilian policymaking. Drawing on the Overton database, we analyze policy documents from federal institutions to trace the incorporation of research outputs, highlight the role of key funding agencies, and assess trends in the reliance on domestic versus international scientific evidence.

Method

This study utilizes the Overton policy database to analyze the integration of scientific research into Brazilian policy documents. The Overton database, established through web-crawling publicly accessible documents from over 43,000 organizations found in more than 2,000 policy sources, is one of the world's largest repositories of policy documents, encompassing governments, intergovernmental organizations (IGOs), think tanks, and charitable entities. As of May 2024, the database included more than 13 million policy documents. Each document in Overton is processed to extract bibliographic information such as title, authors, and publication date, along with cited references from academic literature and other policy documents. Overton's broad definition of policy documents includes materials primarily written for or by policymakers, such as reports, clinical guidelines, white papers, and legal manuscripts. The database's coverage spans documents from almost 200 countries written in several languages.

We conducted a search within Overton for sources in the Brazilian government. In total, Overton indexed 24 organizations in 12 policy sources in Brazil. Of those, seven are from the Brazilian government, while the others are Brazilian IGOs, and I think thanks. The search yielded 109,769 policy documents published between 1997 and 2023. We exported 60.458 unique Digital Object Identifiers (DOIs) from research cited in these policy documents and searched them in the Web of Science (WoS) database. The results yielded 35,230 research documents (58,3%), and the metadata from these documents was later exported to the VantagePoint software for

further processing and analysis. Data was cleaned and standardized, focusing on countries, organizations, journals, and funding agencies.

Results and discussion

Results from cited research in Brazilian government policy documents show that 95.5% were written in English and only 4.1% in Portuguese. Additionally, 83.4% are articles, 7% are reviews, and 5% are proceeding papers. Figure 1 shows the countries with at least 400 cited research documents in the sampled policy documents, with Brazil highlighted in black. The analysis of scientific citations in Brazilian policy documents reveals a diverse array of international influences, with the United States leading significantly. The United States' research, cited 13,994 times, indicates a substantial reliance on American scientific output, reflecting the country's global leadership in various research fields. Brazil itself ranks second with 6,350 citations, underscoring the weight of domestic research in policy formulation. The prominence of Brazilian research in policy documents indicates its relevance to certain policy areas, particularly in health and environmental sciences. These fields also benefit from the support of Brazil's well-established research institutions, such as Fiocruz, Embrapa, and INPE, which play pivotal roles in these domains as previously cited.

However, the extent to which the academic community's research agenda aligns with broader governmental priorities remains uncertain and likely varies across different fields. While some topics may reflect policy needs, others may be more influenced by academic interests, funding availability, or international research trends. Nonetheless, from a public policy perspective, the presence of research authored by Brazilian scholars is significant, as it increases the likelihood that locally relevant knowledge, methodologies, and contexts are considered in policy formulation.

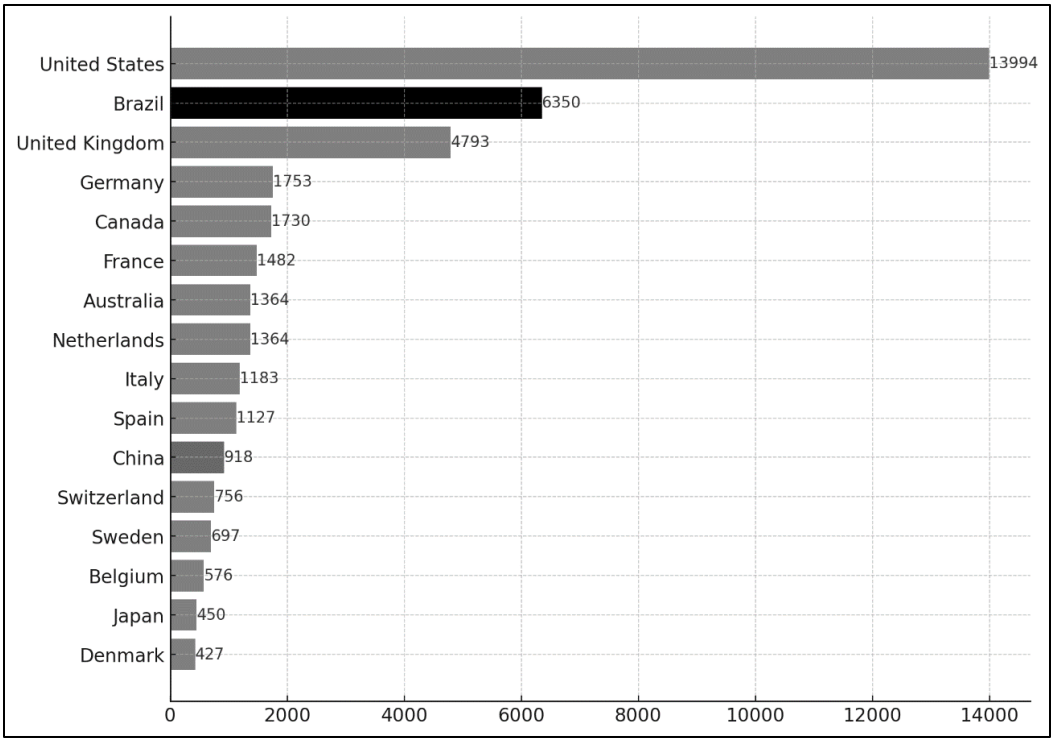


Figure 1. Most cited countries in sampled policy documents.

The dominance of the Global North is evident, with countries like the United Kingdom, Germany, Canada, and others contributing substantially. This reflects a broad spectrum of scientific collaboration and influence from developed nations. China, with 918 citations, is the only Global South country among the top 16 cited countries, highlighting a significant disparity in the sources of scientific research. Beyond China, other Global South countries such as Argentina (357 citations), India (331 citations), South Africa (312 citations), Mexico (275 citations), and Chile (218 citations) have a notable but comparatively smaller presence. This pattern underscores the dominance of research from the Global North in Brazilian policy documents while still recognizing the valuable contributions from a few key Global South nations. This aligns with previous findings that highlight the challenges faced by Global South nations in bridging the science-policy gap due to limited international collaboration (Szomszor & Adie, 2022).

The funding landscape further underscores the challenges and contributions of national and international agencies. Figure 2 highlights the pivotal roles of domestic funding bodies like the National Council for Scientific and Technological Development (CNPq) and the São Paulo Research Foundation (FAPESP). Despite significant financial constraints, FAPESP has remained resilient, maintaining its state-mandated funding to support critical research projects (Faleiros, 2018).

International funding agencies also play a significant role, with the National Institutes of Health (NIH) and the National Science Foundation (NSF) of the United States being major contributors. These figures illustrate the influence of U.S. funding

on Brazilian research outputs and its integration into policy frameworks. The Economic and Social Research Council (ESRC) from the United Kingdom and the European Union (EU) are other key international contributors. The presence of the National Natural Science Foundation of China (NSFC) highlights China's growing influence in global research collaborations.

In addition to FAPESP, other Brazilian regional funding agencies, such as the Rio de Janeiro Research Foundation (FAPERJ) and the Minas Gerais Research Foundation (FAPEMIG), showcase the significant contributions of state-level funding to the national research ecosystem. Additionally, specialized institutions like the Brazilian Synchrotron Light National Laboratory (LNLS) reflect the impact of targeted research infrastructure investments.

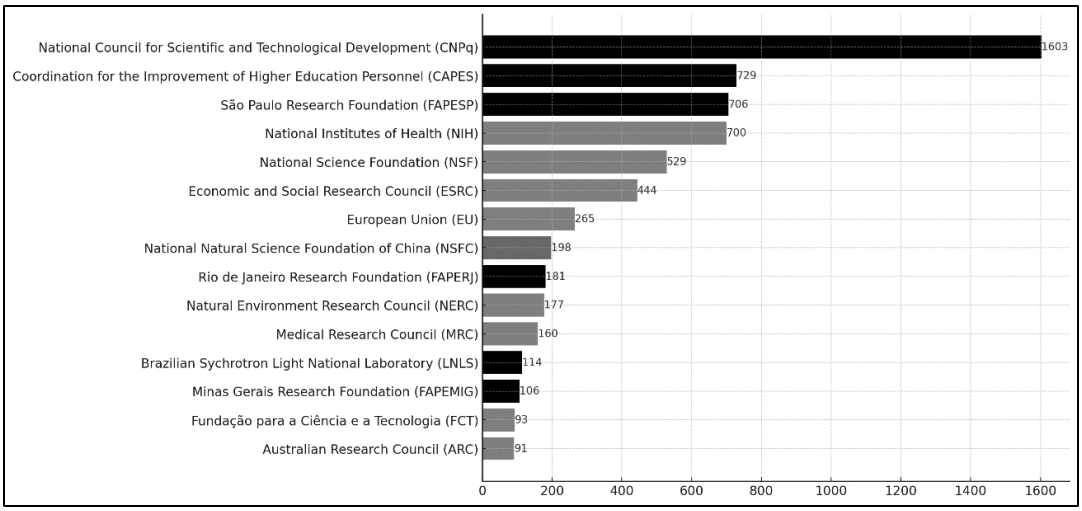


Figure 2. Main funders of cited research in sampled policy documents.

Domestic institutions also feature prominently in policy citations, as shown in Figure 3. The University of São Paulo (USP), the State University of Campinas (UNICAMP), and the Federal University of Rio de Janeiro (UFRJ) lead as key contributors to policy-relevant research. This reflects the ability of Brazilian institutions to align their research outputs with national policy priorities despite systemic underfunding (Escobar, 2022). However, the influence of research on policy is not solely determined by the volume of citations but also by the relevance and accessibility of the research outputs.

Internationally, the National Bureau of Economic Research (NBER) and Harvard University are also highly influential. These institutions, along with the University of Oxford, Stanford University, and the University of London, contribute significantly to the research base that Brazilian policymakers draw upon. This indicates a strong reliance on leading global academic and research institutions to support policy decisions in Brazil.

Beyond other Brazilian universities and research organizations, there are not many Global South organizations frequently cited. The Chinese Academy of Sciences

(CAS) and the University of Cape Town (UCT) are the most cited ones, but with only 98 and 84 citations, respectively. On the other hand, the inclusion of other Global North institutions demonstrates the breadth of international research informing Brazilian policies.

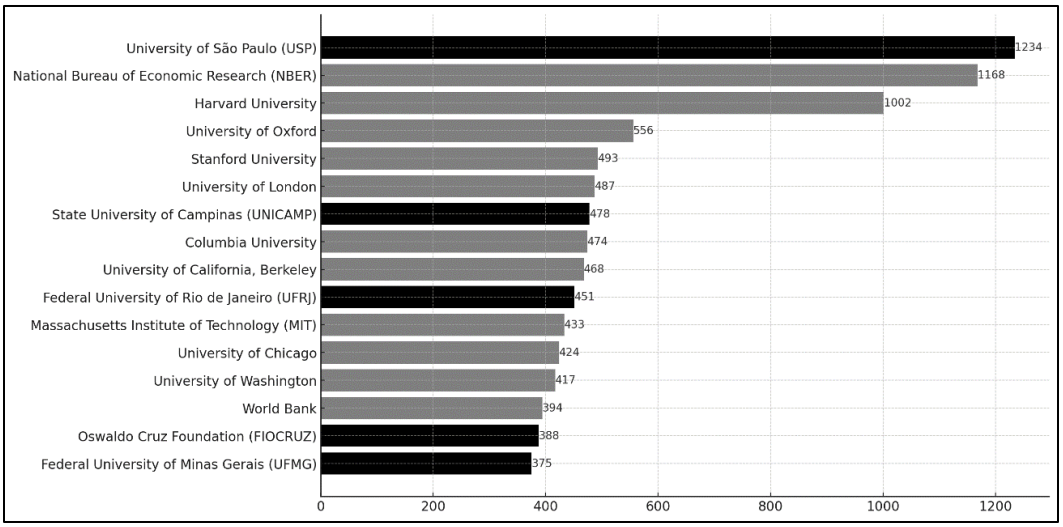


Figure 3. Organizations with the most cited research in sampled policy documents.

The analysis of the most cited journals in Brazilian policy documents and their impact factor reveals a diverse range of influential publications across various fields (Figure 4). Leading the list is the American Economic Review with 450 citations, followed by the Journal of Political Economy, the Quarterly Journal of Economics, and a range of other economics journals, highlighting the strong influence of economic research on Brazilian policymaking.

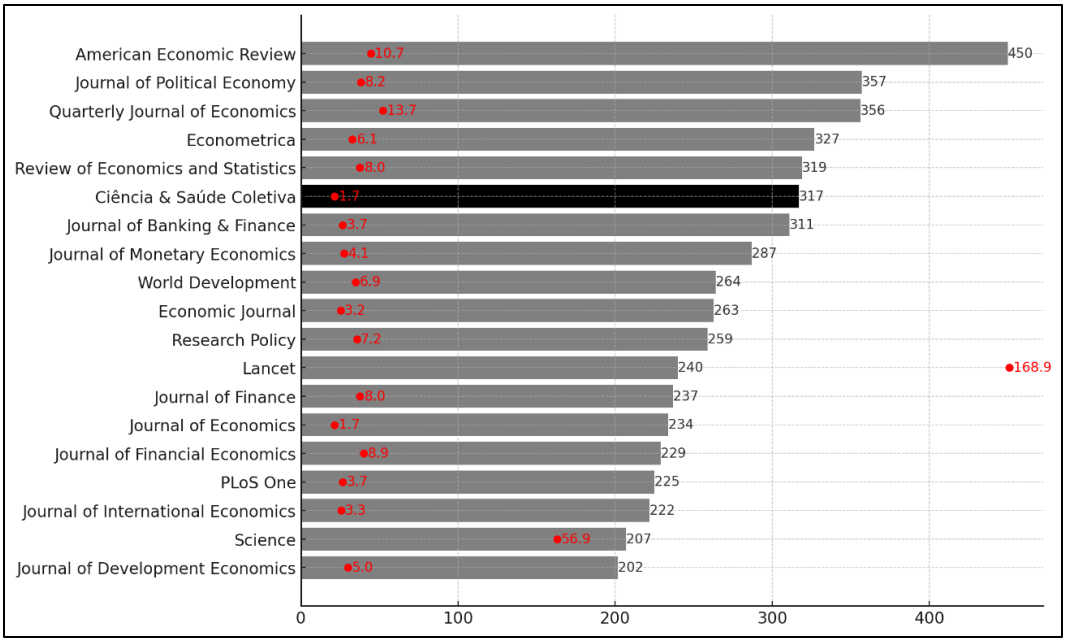


Figure 4. Journals with the most cited research in sampled policy documents.

Ciência e Saúde Coletiva stands out as the sole Brazilian journal with significant influence, accumulating 317 citations. This reflects the journal's critical role in disseminating health-related research that informs public health policies in Brazil. Despite its relatively lower impact factor of 1.7, its frequent citation underscores the practical relevance and impact of its published research on Brazilian public health policy.

Other highly cited journals include international publications like the Lancet, which, despite having a high impact factor of 168.9, indicates the integration of high-impact global health research into Brazilian policy frameworks. Similarly, Science, with an impact factor of 56.9, and PLoS One, with a more moderate impact factor, illustrate the broad scope of scientific research considered in Brazilian policymaking.

Conclusion

Despite significant financial constraints and the challenges posed by science denialism and political narratives in recent years, Brazilian science remains a relevant force in shaping evidence-based public policies. Domestic research institutions play a relevant role in aligning scientific outputs with national priorities, particularly in health and environmental management. These institutions are supported by national and state-level funding agencies that help sustain scientific production. However, the analysis also reveals Brazil's dependency on Global North research, reflecting structural inequities in global knowledge production. To strengthen the science-policy interface, Brazil must prioritize investments in research infrastructure, enhance South-South collaborations, and bolster mechanisms that promote the visibility and accessibility of domestic research. Such efforts are

essential not only to safeguard Brazil's scientific legacy but also to ensure its long-term contribution to global and national policymaking.

References

Angelo, C. (2016). Brazil's scientists battle to escape 20-year funding freeze. *Nature*, *539*(7630), 480–480. https://doi.org/10.1038/nature.2016.21014

Angelo, C. (2017). Brazilian scientists reeling as federal funds slashed by nearly half. *Nature*. https://doi.org/10.1038/nature.2017.21766

Diele-Viegas, L. M., Hipólito, J., & Ferrante, L. (2021). Scientific denialism threatens Brazil. *Science*, *374*(6570), 948–949. https://doi.org/10.1126/science.abm9933

Escobar, H. (2022). Brazil's science budget is rebounding. So why aren't scientists celebrating? In *Science*. https://doi.org/10.1126/science.ada0660

Faleiros, G. (2018). How science supports São Paulo. *Nature*, *563*(7733), S179–S181. https://doi.org/10.1038/d41586-018-07536-1

Gibney, E. (2015). Brazilian science paralysed by economic slump. *Nature*, *526*(7571), 16–17. https://doi.org/10.1038/526016a

Moutinho, S. (2022). Brazil's election is a cliffhanger for scientists. *Science*, *378*(6617), 235–236. https://doi.org/10.1126/science.adf3946

Szomszor, M., & Adie, E. (2022). Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, *3*(3), 624–650. https://doi.org/10.1162/qss_a_00204

https://doi.org/10.51408/issi2025_117

Not All 'Predators' are the Same: Exploring the Spectrum of Questionable Journals

Zehra Taşkın[1], Güleda Doğan[2], İdris Semih Kaya[3], Ezgi Uğurlu[4], Özge Söylemez[5], Ceren Bilge Seferoğlu[6], Emanuel Kulczycki[7]

[1] *ztaskin@hacettepe.edu.tr*, [2]*gduzyol@hacettepe.edu.tr*
Hacettepe University, Department of Information Management, Ankara (Turkey)

[3] *idosfer@gmail.com*
Hacettepe University, Department of Information Management, Ankara (Turkey)
Ministry of Culture and Tourism, Düzce Provincial Public Library, Düzce (Turkey)

[4]*ezgi.ugurlu@bilkent.edu.tr*
Hacettepe University, Department of Information Management, Ankara (Turkey)
Bilkent University Library, Ankara (Turkey)

[5] *osoylemez@ankara.edu.tr*
Hacettepe University, Department of Information Management, Ankara (Turkey)
Ankara University Library, Ankara (Turkey)

[6] *cseferoglu@bartin.edu.tr*
Hacettepe University, Department of Information Management, Ankara (Turkey)
Bartın University, Department of Information and Records Management, Bartın (Turkey)

[7] *emek@amu.edu.pl*
Adam Mickiewicz University, Scholarly Communication Research Group, Poznań (Poland)

Abstract

So-called 'predatory' publishing is often framed as an issue of unethical journal practices, but this perspective overlooks deeper structural problems in scholarly communication. The reliance on blacklists as a primary solution to identifying questionable journals fails to acknowledge the complexity of academic publishing and the broader systemic issues that contribute to unethical or controversial publishing practices. These include not only so-called 'predatory' journals but also concerns such as 'special issue-ization' and the rise of paper mills. Furthermore, the strategies used by emerging open-access mega-publishers increasingly resemble those employed by traditional and hybrid publishers, demonstrating that questionable practices are not confined to a single category of journals. This research in progress critically examines the characteristics of journals labeled as so-called 'predatory' and questions the effectiveness of static blacklists in scholarly assessment. Using a dataset of 2,755 journals from Predatory Reports, we systematically analyze their ISSN registration, subject classifications, accessibility, financial models, editorial transparency, and indexing status. While we recognize the limitations of blacklists, this dataset provides a basis for exploring broader patterns in academic publishing. Preliminary findings reveal that 24% of the journals became inaccessible after being listed, suggesting that some publishers shut down or rebrand to evade scrutiny. While ISSN registration is not mandatory, 13% of the journals in the dataset do not have one, which may indicate variations in registration practices. The geographical distribution of these journals is concentrated in India (31.45%), Switzerland (30.17%), and the United States (21.36%). This distribution highlights the global nature of these practices, spanning a range of publication models. The study also finds that 71% of these journals charge Article Processing Charges (APCs), while 23.7% fail to disclose APCs before submission, creating financial uncertainty for authors.

Rather than indiscriminately covering all fields, many journals now focus on STEM disciplines. These findings underscore the need for more nuanced, criteria-based evaluation frameworks that account for the complexities of scholarly publishing, moving beyond binary categorizations of journals as 'predatory' or legitimate.

Introduction

The phenomenon of so-called 'predatory' publishing is often portrayed as a pressing concern in academic research, but its implications extend beyond exploitative practices by questionable publishers. At its core, the issue reflects deeper systemic inequalities in scholarly communication, where access to resources and opportunities for publishing high-quality research are unevenly distributed (Krawczyk & Kulczycki, 2021; Kulczycki, 2023). While so-called 'predatory' publishers exploit the open-access model for financial gain, bypassing quality control and undermining trust in academic outputs (Grudniewicz et al., 2019), framing the problem purely in terms of "predators" and "victims" oversimplifies a much more complex issue.

Labeling journals as so-called 'predatory' or legitimate creates binary categorizations that fail to account for the diversity of practices among questionable publishers and the systemic issues driving these dynamics. This approach, often operationalized through blacklists, has significant limitations. Blacklists are difficult to maintain and update, particularly when dealing with journals backed by powerful commercial interests (Ryan, 2024; Silver, 2017). While open-access mega publishers have often been scrutinized for lapses in quality control (Fränti, 2024; Mills et al., 2024; Oviedo-García, 2021), recent research suggests that commercially driven publishing strategies extend beyond these newer models and are also present in traditional and hybrid publishers (Shu & Larivière, 2024). The broader challenge is not exclusive to a specific type of publisher but rather reflects evolving strategies across the scholarly publishing landscape (Nicholas et al., 2023). The debate surrounding these issues further highlights the limitations of a binary framework, as publishing practices increasingly defy simple categorization (Tsigaris & Teixeira da Silva, 2021).

Moreover, reliance on blacklists perpetuates inequities in research evaluation by prioritizing the journal's reputation and indexing status over the actual content or contributions of the research itself. This is particularly evident in research assessment systems that use journal-based metrics as proxies for scholarly quality, influencing hiring, funding, and promotion decisions (Mills & Inouye, 2021; Öztürk & Taşkın, 2024). In peripheral academic contexts, where scholars may face additional barriers to publishing in high-impact journals, these pressures push researchers toward venues that may later be labeled as questionable. Rather than reflecting individual choices alone, such publishing patterns often stem from structural inequalities within global academia (Mertkan et al., 2021; Taşkın et al., 2023).

Further complicating the landscape, large language models (LLMs) like ChatGPT introduce new challenges for academic publishing. These tools enable the rapid generation of text, which has already been exploited to produce papers for paper mills, amplifying unethical publishing practices (Kendall & Teixeira da Silva, 2024). However, LLMs are not the root cause of these issues. The exponential growth of the publish-or-perish culture, driven by quantity-focused research evaluation

systems, has created an environment where such technologies can flourish. While LLMs are positioned as a new scapegoat, the real challenge lies in addressing the systemic pressures that prioritize publication quantity over quality. Policymakers, editors, and publishers must develop strategies not only to mitigate the misuse of LLMs but also to reform evaluation systems that perpetuate these issues, ensuring that scholarly communication prioritizes meaningful contributions over sheer output. This research-in-progress does not aim to classify journals as 'predatory' or legitimate but instead critically examines the broader risks of such dichotomies. By analyzing factors such as accessibility, publication origins, subject categories, languages, and editorial practices, this study seeks to highlight the structural issues that contribute to so-called 'predatory' publishing. Ultimately, the goal is to inform policies that shift the focus from where research is published to the societal and scientific contributions it makes, promoting responsible and equitable research evaluation practices.

Methods

For this study, we utilized the list of so-called 'predatory' journals available on the Predatory Reports website, [1] as it represents one of the most extensive and up-to-date resources accessible. Despite the anonymity of its creators,[2] which is understandable given the challenges faced by earlier efforts in this field (Ryan, 2024; Silver, 2017), the list was selected for its broad scope and inclusion of diverse journal types. This allowed us to create a large dataset for in-depth examination.

While we do not endorse blacklists as a definitive tool for evaluating journal quality, we use this dataset as a starting point to analyze broader publishing patterns. Rather than assuming the journals listed are inherently unreliable, we examine their characteristics systematically to understand the diverse operational models that lead to their inclusion. Our study builds on previous research by focusing not only on journal attributes but also on their accessibility, financial practices, and indexing status over time, offering insights into how such classifications evolve.

To achieve this, we analyzed a dataset of 2,755 journals from the Predatory Reports list. The journals are systematically evaluated across multiple dimensions, and detailed data about their practices is being collected. The data collection focuses on key aspects of journal operations, grouped into the following categories:

- **Identification and registration:** This includes verifying whether the journal is registered in the ISSN portal, the associated country, and the accuracy of provided ISSN information.
- **Website and accessibility:** The website language, availability of an English version, clarity of scope (e.g., interdisciplinary or specific fields), and access to full-text articles or metadata are assessed.

[1] The list was downloaded on 12 September 2024 from https://predatoryjournals.org/predatory-journals.

[2] It is indicated in the website as: "We decided to remain anonymous so as not to be sued by companies whose practices are quite aggressive. As our reach grows on the internet, we are already receiving threats."

- **Financial practices:** The presence and transparency of APCs, including the cost, discount options, and details on how funds are utilized.
- **Editorial and peer review processes:** This involves checking for information on editorial boards, peer review processes, and guarantees of publication timelines (e.g., fast publication promises).
- **Indexing and metrics:** The journal's indexing status in citation indexes and bibliographic databases, along with the inclusion of citation metrics and their sources, are documented.
- **Publishing policies and licensing:** The presence of licensing policies (e.g., Creative Commons) and details on publishing rights and practices are recorded.
- **Transparency and contact information:** The availability of publisher contact details, such as email, phone, and physical addresses, as well as the credibility of listed editorial and reviewer boards.

This research is ongoing, and the collected data provides a foundation for understanding the diverse characteristics of these journals. In this paper, we present preliminary findings on the availability of journal websites, their geographic distribution, field distribution, and APC transparency. Further analysis of editorial practices, indexing status, and licensing policies will be conducted in future stages of this research.

Preliminary findings

Our initial analysis revealed that 24% of the journals listed on the Predatory Reports platform became inaccessible following their inclusion in the list. This finding highlights a significant issue: some journals associated with questionable practices often shut down their operations, remove published articles, or rebrand under a different name as soon as they are labeled. Additionally, we found that 13% of these journals are not registered in the ISSN portal, which may reflect variations in registration practices rather than a definitive indicator of legitimacy. The absence of an ISSN complicates efforts to track and evaluate these journals over time. Figure 1 shows the geographic distribution of journals with ISSN registration.

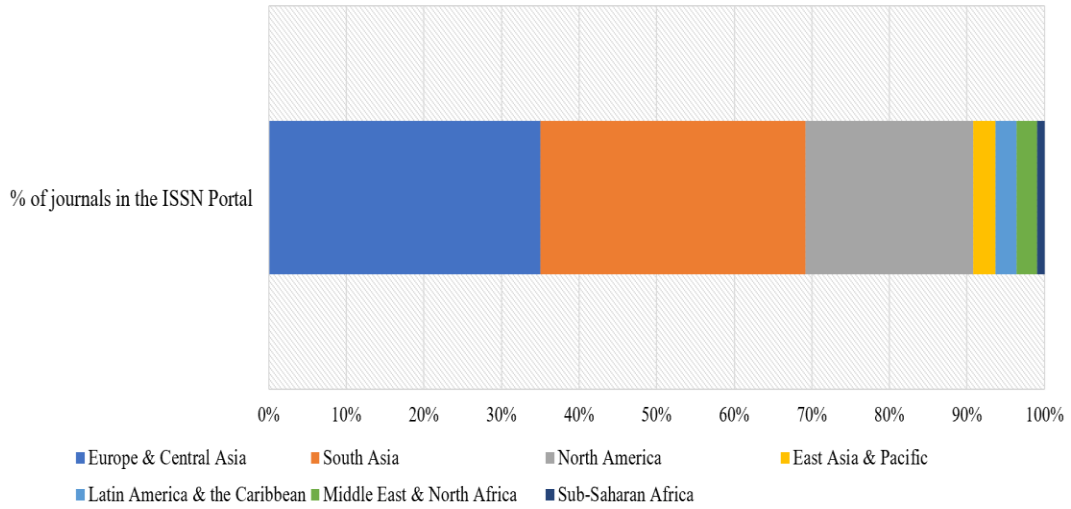


Figure 1. Geographic distribution of journals labeled as predatory that are registered in the ISSN Portal.

The distribution of journals labeled as so-called 'predatory' by country highlights significant global patterns in academic publishing. India accounts for the highest proportion of these journals (31.45%), followed by Switzerland (30.17%) and the United States (21.36%). Other notable contributors include Brazil (2.26%), Pakistan (1.57%), Turkey (1.53%), and Iran (1.03%). This distribution suggests that so-called 'predatory' publishing is not solely an issue of individual journal practices but is influenced by broader systemic and geopolitical factors. However, it is important to acknowledge the limitations of the dataset, as lists such as Predatory Reports or Beall's list tend to focus on journals from peripheral academic systems and may not comprehensively capture journals operating within more established publishing networks (Krawczyk & Kulczycki, 2021).

The classification of journals by subject fields was based on their own declarations on their websites. Journals that explicitly indicated their focus on science, technology, engineering, and mathematics (STEM) were categorized accordingly, while those emphasizing social sciences and humanities (SSH) were grouped separately. Journals that described themselves as interdisciplinary or covering multiple broad areas (e.g., sciences, social sciences, and humanities) were categorized as interdisciplinary. Additionally, journals that did not provide any subject classification on their websites were labeled as "unknown."

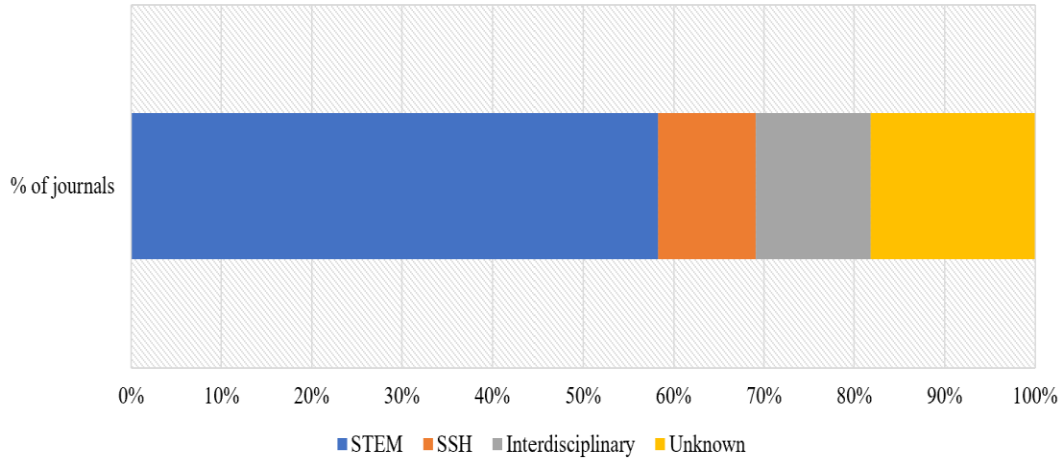


Figure 2. Field distribution of journals labeled as predatory.

The initial perception of so-called 'predatory' publishing is that it indiscriminately covers all fields without clear specialization. However, our data suggests that many of these journals now indicate specific subject areas for publication. This shift may reflect a response to increased scrutiny of broad-scope journals. However, prior research has highlighted that one questionable practice associated with these journals is publishing out-of-scope papers. Thus, while some journals claim subject specialization, this does not necessarily equate to maintaining field-specific editorial standards.

The analysis of APCs among journals labeled as so-called 'predatory' reveals that 71% require APCs, reinforcing the notion that financial gain can be a primary driver of their operations. However, 23.7% provide no clear information about APCs before submission, creating uncertainty for authors who may only learn about the charges after their manuscripts have been accepted. This lack of transparency is a key indicator of deceptive publishing practices. Interestingly, only 5% explicitly state that they do not require APCs, while an even smaller subset (0.08%) request payments as "donations." Future analysis will examine whether journals disclose how APC revenues are allocated, if such information is available on their websites.

Future steps

This research-in-progress has presented preliminary findings on ISSN registration and field classifications. Moving forward, we will expand the analysis to other key dimensions to provide a more comprehensive understanding of so-called 'predatory' journals.

One priority is tracking website availability over time to determine whether journals rebrand or disappear, suggesting adaptive strategies. We will also investigate financial transparency, focusing on APC disclosure and potential hidden costs, with the hypothesis that unclear APC policies contribute to author exploitation.

Editorial and peer review practices will be examined to assess transparency in editorial boards and peer review claims, particularly regarding fast-track publication

promises. Additionally, we will analyze indexing and citation metrics to verify impact claims and assess how these journals establish credibility.

Finally, we plan to study licensing and archiving policies to determine whether these journals ensure long-term access to published work. These ongoing analyses will contribute to a more nuanced understanding of questionable publishing practices, informing responsible research evaluation frameworks.

Conclusion

Our project moves beyond binary classifications of so-called 'predatory' journals to provide a more nuanced understanding of questionable publishing practices. Instead of relying on static blacklists, it systematically examines journal operations, financial models, indexing claims, and accessibility to reveal broader trends in scholarly publishing.

Beyond academic publishing, the findings inform research policy by promoting more transparent and responsible evaluation frameworks. By analyzing accessibility, editorial transparency, and financial disclosures, the project helps institutions, funding bodies, and scholars make more informed decisions, shifting the focus from journal labels to the quality and impact of research.

Acknowledgments

This study was conducted as part of the project titled "Preventing Questionable Publishing Practices through Responsible Research Evaluation Policies that Consider Geopolitical Dynamics", funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK), grant number 223K471.

Author contributions

ZT: Conceptualization, methodology, data analysis, visualization, supervision, writing-original draft | GD: Conceptualization, methodology, writing-review & editing | İSK, EU, ÖS, CBS: Methodology, data curation | EK: Conceptualization, methodology, supervision, writing-review & editing

References

Fränti, P. (2024). *What is wrong with MDPI: Is it a predator or a serious competitor?* (arXiv: 2411.08051). arXiv. https://doi.org/10.48550/arXiv.2411.08051

Kendall, G., & Teixeira da Silva, J. A. (2024). Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing, 37*(1), 55–62. https://doi.org/10.1002/leap.1578

Krawczyk, F., & Kulczycki, E. (2021). On the geopolitics of academic publishing: The mislocated centers of scholarly communication. *Tapuya: Latin American Science, Technology and Society, 4*(1), 1984641. https://doi.org/10.1080/25729861.2021.1984641

Kulczycki, E. (2023). *The Evaluation Game: How Publication Metrics Shape Scholarly Communication*. Cambridge University Press. https://doi.org/10.1017/9781009351218

Mertkan, S., Onurkan Aliusta, G., & Suphi, N. (2021). Profile of authors publishing in 'predatory' journals and causal factors behind their decision: A systematic review. *Research Evaluation, 30*(4), 470–483. https://doi.org/10.1093/reseval/rvab032

Mills, D., & Inouye, K. (2021). Problematizing 'predatory publishing': A systematic review of factors shaping publishing motives, decisions, and experiences. *Learned Publishing*, *34*(2), 89–104. https://doi.org/10.1002/leap.1325

Mills, D., Mertkan, S., & Onurkan Aliusta, G. (2024). 'Special issue-ization' as a growth and revenue strategy: Reproduction by the "big five" and the risks for research integrity. *Accountability in Research*, 1–19. https://doi.org/10.1080/08989621.2024.2374567

Nicholas, D., Herman, E., Abrizah, A., Rodrí-guez-Bravo, B., Boukacem-Zeghmouri, C., Watkinson, A., ÅšwigoÅ„, M., Xu, J., Jamali, H. R., & Tenopir, C. (2023). Never mind predatory publishers"¦ what about "˜grey´ publishers? *Profesional de La Información / Information Professional*, *32*(5), Article 5. https://doi.org/10.3145/epi.2023.sep.09

Oviedo-García, M. Á. (2021). Journal citation reports and the definition of a predatory journal: The case of the Multidisciplinary Digital Publishing Institute (MDPI). *Research Evaluation*, *30*(3), 405–419a. https://doi.org/10.1093/reseval/rvab020

Öztürk, O., & Taşkın, Z. (2024). How metric-based performance evaluation systems fuel the growth of questionable publications? *Scientometrics*, *129*(5), 2729–2748. https://doi.org/10.1007/s11192-024-04991-8

Ryan, J. (2024). Exposing predatory journals: Anonymous sleuthing account goes public. *Nature*. https://doi.org/10.1038/d41586-024-03321-5

Shu, F., & Larivière, V. (2024). The oligopoly of open access publishing. *Scientometrics*, *129*(1), 519–536. https://doi.org/10.1007/s11192-023-04876-2

Silver, A. (2017). Controversial website that lists 'predatory' publishers shuts down. *Nature*. https://doi.org/10.1038/nature.2017.21328

Taşkın, Z., Krawczyk, F., & Kulczycki, E. (2023). Are papers published in predatory journals worthless? A geopolitical dimension revealed by content-based analysis of citations. *Quantitative Science Studies*, *4*(1), 44–67. https://doi.org/10.1162/qss_a_00242

Tsigaris, P., & Teixeira da Silva, J. A. (2021). Why blacklists are not reliable: A theoretical framework. *Journal of Academic Librarianship*, *47*(1), 102266. https://doi.org/10.1016/j.acalib.2020.102266

https://doi.org/10.51408/issi2025_127

Old but Not Obsolete: Bag-of-Words vs. Embeddings in Topic Modeling

Jean-Charles Lamirel[1], Francis Lareau[2], Christophe Malaterre[3]

[1] *jean-charles.lamirel@loria.fr*
Université de Strasbourg, SYNALP-LORIA,
615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France (France)

[2] *lareau.francis@courrier.uqam.ca*
Université de Sherbrooke, Dept of Philosophy,
2500, boul. de l'Université, Sherbrooke (QC) J1K 2R1 (Canada),
Université du Québec à Montréal, Dept of Philosophy & CIRST,
455 bd. René-Lévesque Est, Montréal (QC) H3C 3P8 (Canada)

[3] *malaterre.christophe@uqam.ca*
Université du Québec à Montréal, Dept of Philosophy & CIRST,
455 bd. René-Lévesque Est, Montréal (QC) H3C 3P8 (Canada)

Abstract

Topic modeling techniques, including classical Bag-of-Words (BOW)-based methods like Latent Dirichlet Allocation (LDA) and emerging embedding-based models such as Top2Vec and BERTopic, are pivotal for uncovering latent themes in text corpora. This study builds upon previous work on an alternative BOW-based approach relying on feature maximization, CFMf, addressing limitations and extending comparisons along multiple metrics. Using a corpus of philosophy of science research articles (*N*=16,917), we evaluate LDA, CFMf, Top2Vec, and BERTopic across coherence, diversity, and recall metrics while also qualitatively examining top-word interpretability. Results reveal distinct trade-offs: while Top2Vec excels in coherence and diversity, it underperforms in recall and interpretability; BERTopic marginally outperforms LDA in coherence but not recall; CFMf balances these dimensions, outperforming others in coherence and diversity. Findings highlight the enduring relevance of BOW-based models and emphasize the modularity of topic modeling pipelines, advocating for hybrid approaches that integrate optimal components for improved performance.

Introduction

Topic modeling is a cornerstone in computational text analysis, aiming to uncover hidden themes in large corpora. Classical approaches, such as Latent Dirichlet Allocation (LDA), rely on statistical methods based on the Bag-of-Words (BOW) representation. Recently, embedding-based models such as Top2Vec and BERTopic have emerged as promising alternatives. In prior research, we highlighted the performance of a novel BOW-based method, Clustering and Feature Maximization with F1-measure (CFMf), though limitations remained, notably the generation of marginal topics with high document counts (Lamirel et al., 2024). The present study builds upon this work by addressing three objectives. First, we aim to mitigate the residual defects of CFMf. Second, we extend our comparative framework to include transformer-based models like BERTopic, which leverage Large Language Models (LLMs) and long-text embeddings. Finally, we investigate the modular nature of topic modeling, hypothesizing that combining the best components of various

approaches may yield a hybrid, high-performing model. Using a corpus of 16,917 philosophy of science research articles, we evaluate LDA, CFMf, Top2Vec, and BERTopic across multiple performance metrics, including coherence, diversity, and recall measures.

Methods overview

Topic models rely on a broad range of approaches to reveal hidden themes in extensive text corpora. Focusing on LDA, CFMf, Top2Vec, and BERTopic, we briefly describe these approaches notably in terms of text preprocessing, vectorization, clustering, ranking of documents and of words.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a generative statistical model that considers each document as a mixture of topics, each being a mixture of words with specific probabilities. It involves estimating Dirichlet distributions using techniques like Gibbs sampling or variational inference. It starts with tokenization, converting documents into word tokens, then representing them as Bag-Of-Words (BOW) vectors that quantify the tokens in each document. LDA's probabilistic clustering enables ranking of documents and words.

CFMf combines Feature Maximization (Lamirel et al., 2016) for feature selection, based on the F-measure, and Growing Neural Gas (GNG) for neural clustering (Fritzke, 1994). GNG is a winner-take-most algorithm that can utilize various metrics to capture a dataset's topology. To address a text size clustering bias observed when using the classical Euclidean metric (Lamirel et al., 2024), an angular metric is now deployed by renormalizing the cluster's prototype vectors during each learning step. GNG, like LDA, requires the number of topics beforehand. Key steps involve tokenization and BOW vectorization with a normalized TFIDF scheme. GNG clusters documents into topics, while the F-measure ranks words within topics. Cosine distance between topic's prototypes and documents is used for ranking documents.

Top2Vec original model (Angelov, 2020) utilizes Doc2Vec (Le & Mikolov, 2014) for semantic embedding of words and documents. Using HDBSCAN clustering technique (Campello et al., 2013), dense clusters emerge based on data density without the need to specify the number of topics. Each cluster is represented by its centroid taken as the average of cluster document embeddings. Top2Vec considers clusters as topics, using cosine similarity to centroids for reassigning ambiguous documents identified by HDBSAN. Key steps include tokenization and word/document embedding representation.

BERTopic original model (Grootendorst, 2022) employs transformer models like BERT (Devlin et al., 2019) to create deep contextual embeddings. HDBSCAN clusters documents using these embeddings. A BOW representation is used to rank words and documents through class-based TFIDF scores (c-TFIDF). Ambiguous cases from HDBSCAN are reassigned via cosine similarity between c-TFIDF representations. The process entails tokenization and dual vector representation: transformer-based for clustering and BOW-based for topic reassignment and document/word ranking.

Experimental protocol

The dataset comprised the complete collection of 16,917 full-text research articles from eight leading philosophy of science journals, as curated by Malaterre and Lareau (2022) and covering the period from 1930 to 2017. The corpus underwent standard preprocessing steps: tokenization, part-of-speech tagging, and lemmatization (TreeTagger package (Schmid, 1994) with Penn TreeBank (Marcus et al., 1993)). Words appearing in fewer than 50 sentences were excluded; only nouns, verbs, adverbs, and adjectives were retained. Documents were then vectorized to produce term-document matrices (TDMs) based on word frequencies for LDA and BERTopic, and on normalized TFIDF for CFMf.

LDA modeling was conducted via a Python API and used a word frequency TDM. CFMf was implemented with custom C and C++ code, using a normalized TFIDF TDM. Top2Vec was executed using a Python API, with the preprocessed corpus transformed by Doc2Vec serving as input. For BERTopic, full-text documents were used as inputs for generating document embeddings through a start-of-the-art method noted for its best average score on the Massive Text Embedding Benchmark Leaderboard: the stella model (stella_en_1.5B_v5) based on Alibaba-NLP and supporting the representation of long texts (131,072 tokens or more). BERTopic standard pipeline was performed with a Python API, using also the TDM for word ranking and outlier reassignment.[1]

Models were built for a number of topics from $K = 5$ to 100. For LDA and CFMf, predetermined values were chosen to sample this interval. For Top2Vec and BERTopic, specific values for the parameter corresponding to minimum cluster size were chosen through trial and error to generate models with different K values. Note that the terms "cluster", "class", or "topic" are used interchangeably. CFMf, Top2Vec, and BERTopic perform crisp clustering of documents and extract top-terms representing topics shared by documents of the same clusters. In contrast, LDA considers documents as probability distributions over topics; crisp clustering is obtained by grouping documents based on their dominant topic.

To compare model performance along complementary dimensions, four measures were used: (i) coherence, which indicates the extent to which top words in each topic are more meaningful when considered together (we used the coherence C_V of Röder et al. (2015)); (ii) topic diversity, which measures the distinctness of topic top words (expressed as the ratio of the number of unique top words in all topics by the total number of top words in all topics); (iii) a measure we call "micro inner recall" (*mIR*) which indicates the extent to which topic top words are found, on average, in the topic documents; and (iv) and "micro joint inner recall" (*mJIR*), which indicates how

[1] LDA Python API: https://github.com/lda-project/lda; CFMf implemented with custom C and C++ code available upon request (plans are to translate this method into Python and transfer it to GitHub in the near future); Top2Vec Python API https://github.com/ddangelov/Top2Vec; Doc2Vec Gensim implementation: https://github.com/piskvorky/gensim; BERTopic API: https://maartengr.github.io/BERTopic/api/bertopic.html; stella model stella_en_1.5B_v5, https://huggingface.co/dunzhang/stella_en_1.5B_v5.

well the top words of the clusters can all jointly recall the documents associated with these clusters. The latter two can be expressed as:

$mIR = \frac{1}{W \times |D|} \sum_{c=1}^{K} \sum_{i \in Top_W[c]} |\{d \in c \mid d[i] \neq 0\}|$ $mJIR = \frac{1}{|D|} \sum_{c=1}^{K} |\{d \in c \mid \exists\, i, i \in Top_W[c] \mid d[i] \neq 0\}|$

where W is the number of top words chosen as description of any cluster c, $|D|$ is the number of documents in the corpus, $Top_W[c]$ is the set of the top W words describing topic c, $d[i]$ represents the presence/absence of word i in the document d.

To gain qualitative insights into the relative topical coverage of the models and facilitate top-word comparison, clusters generated by CFMf, Top2vec and BERTopic were aligned to previously interpreted LDA topics (K=25) based on maximum number of shared documents.

Results

Results of the coherence measures across models show that coherence increases as a function of the number of topics K, reaching some sort of plateau after 50 topics for BERTopic or even earlier around 20-30 topics for the other three models (Fig. 1A). This indicates that topic top words tend to be specific to more narrowly defined clusters as K increases. Of the four approaches, Top2Vec displays the highest coherence at about 0.8 from K=20 onward. CFMf follows with coherence above 0.7 also from K=20 onward. While LDA exhibits the lowest coherence scores, reaching a plateau of about 0.55 from K=20 onward, it is slightly outperformed by BERTopic at lower K values and more significantly at higher K values.

As the number of topics K increases, topic diversity tends to decrease (Fig. 1B), which is to be expected since increasing the number of topics simultaneously increases the likelihood of overlap between top-words. Highest topic diversity—typically above 0.95—is obtained by Top2Vec across all values of K. CFMf ranks second, with diversity measures decreasing from about 0.9 below K=20 to 0.8 after. LDA and BERTopic reach about the same bottom value of about 0.65 after K=30, though LDA outperforms BERTopic for lower K values.

If one were to evaluate the models solely on coherence and diversity, then Top2Vec would come on top. Yet, the two measures of inner recall show a radically different perspective. In terms of micro inner recall—which is the average capability of topic top words to recall their sets of topic documents—, Top2Vec displays by far the lowest scores, below 0.35 for all K values (Fig. 1C). On the other hand, LDA exhibits the highest scores, consistently above 0.8. BERTopic follows, with *mIR* values decreasing from about 0.8 for K=10 to 0.7 for K=100. As for CFMf, it exhibits *mIR* scores that reach a plateau of about 0.6 from K=25 onward.

Measures of joint inner recall, which is the capability by all top words to jointly recall all corpus documents, single out Top2Vec as the least well-performing approach (Fig. 1D). Indeed, while *mJIR* scores for LDA, BERTopic and CFMf all reach about 1, *mJIR* measures for Top2Vec reach a plateau of about 0.9, starting from 0.7 to 0.8 scores for K values below 25. This shows the inability of top-words generated by Top2Vec to recall a remaining fraction (about 10%) of the corpus, even when increasing the number of topics or clusters.

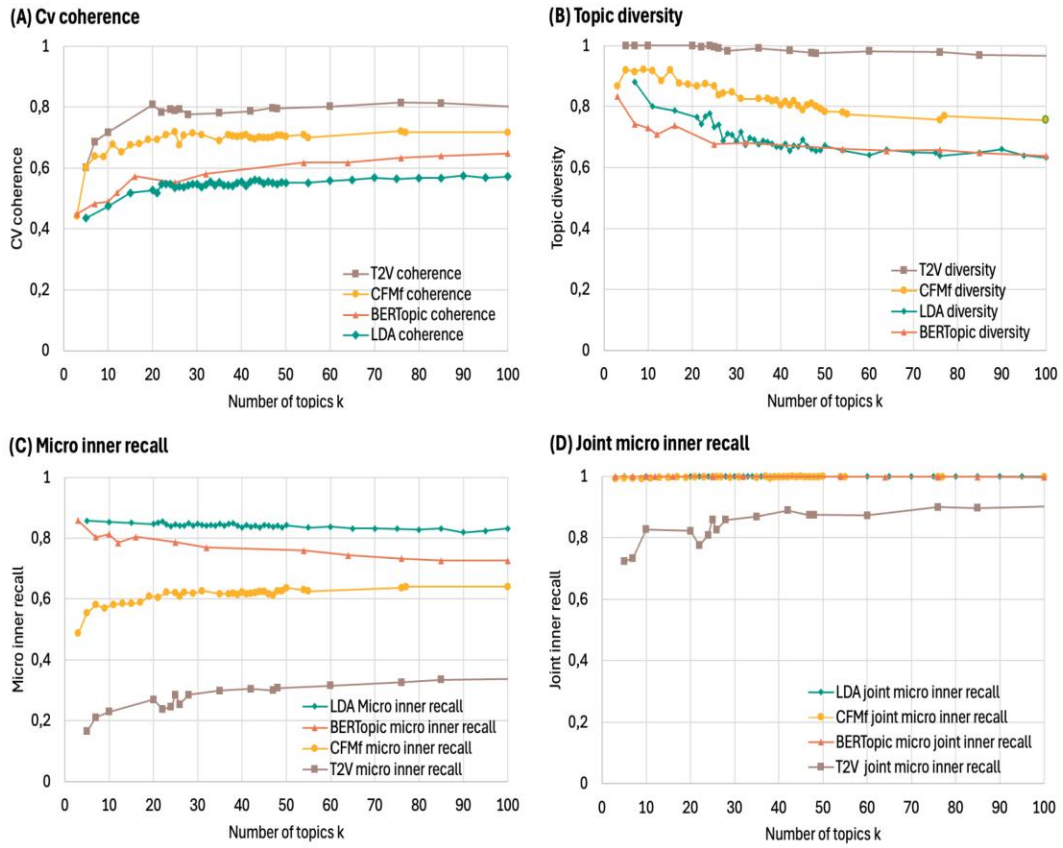


Fig. 1. Performance comparisons between topic models. (A) C_V coherence, (B) Topic diversity, (C) Micro inner recall *mIR*, (D) Micro joint inner recall *mJIR* (for $W = 10$ top-words).

Overall, the four approaches pick out topics that have a good descriptive similarity in terms of top words (Table 1). Yet nuances exist. Most striking is the weaker interpretability of Top2Vec top words, for instance for cluster (21) which mentions author names and technical disciplinary terms, or for cluster (16) which is about causation without naming it but mentioning author names. LDA, CFMf and BERTopic fare better in this respect. While CFMf still mentions author names in some topics—e.g. (8), (17), and (18)—they are fewer than Top2Vec and tend to be well aligned with easily interpretable topics. CFMf top words also tend to convey meaningful interpretations often more precise than LDA, e.g. distinguishing between relativity (2) and quantum mechanics (22), as Top2Vec and BERTopic also do. As for BERTopic, top words are also conducive to clear interpretations, although some of them remain very generic. Note some shifts in the overall balance of topics compared to the LDA model, with fewer topics related to philosophy of language and logic (0, 21, 20) and more topics related to rational decision (14, 18, 19) and especially philosophy of physics (17, 22, 13, 6, 5). What remains to be investigated is whether such changes are also related to changes in the relative proportions of the topics (as expressed in topical percentages or numbers of documents sorted by

dominant topics in LDA or number of cluster documents in BERTopic, CFMf and Top2Vec).

Table 1. Comparison of the top-words for *K*=25. LDA topic colors/labels as in (Malaterre & Lareau, 2022); for CFMf, Top2Vec and BERTopic, colors based on closest LDA topics; numbers are IDs; due to space reasons, only the top 4 words are listed, with abbreviations.

LDA		CFMf		Top2Vec		BERTopic	
Formal	set; function; relation; define.nition	(2)	proba.; propos.; theorem; condition.	(21)	sneed; balzer; moulines	(0)	sentence; truth; language; set
Language	language; sentence; term; meaning	(8)	carnap; wittgenstein; schlick; neurath	(8)	prerequisite; mortgage; exist.; false.	(21)	vague.ness; borderline; predicate
Mathematical	mathematical.tics; number; proof	(6)	math.; axiom; proof; geometry	(24)	nominalistic.ally; indispens.; colyvan	(20)	belief; agent; revision; model
Sentence	sentence; context; use; say	(1)	sentence; language; quine; speaker	(19)	quine.ean; synonymy; gavagai	(7)	belief.ve; know.ledge; epistemic
Truth	logic; truth; true; proposition	(0)	logic; modal; proposition; predicate	(1)	quantifier.cation; provable; semantics	(10)	theory; realist.m; scientific
Arguments	argument; claim; say; question	(9)	belief; epistemic; justif.; doxastic	(20)	ditmarsch; bisimilar; baltag; fagin	(24)	law; nature; generaliz.; statement
Knowledge	belief; knowledge; epistemic; know	(14)	realist.ism; fraassen; putnam	(6)	justified.cation; reliabilist; bivs	(1)	proba.; hypothesis; evidence
Sc.-theory	theory; scientific; empirical; realism	(3)	confirm.ation; hypothesis; inductive	(14)	anti.realist; pessimistic; nma	(14)	game; player; strategy; equilibrium
Confirmation	law; hypothesis; statement; evidence	(11)	proba.; bayes.; frequency; chance	(11)	longino; feminist; kourany; funding	(18)	moral; reason; action; normative
Experiment	datum; experiment; value; use	(10)	agent; game; player; utility	(0)	bayes.ians; probability; finetti	(19)	economic.s; theory; price
Probability	probability; measure; value; give	(24)	selection; pop.; fitness; evolutionary	(18)	morally.ty; baier; utilitarian.ism	(4)	selection; gene; organism; pop.
Agent-	agent; action; decision; game	(23)	gene; cell; organism; protein	(23)	replicat.; payoff; huttegger; signalers	(15)	function; teleological; artefact; goal
Evolution	selection; pop.; organism; gene	(12)	brain; cognitive; machine; mental	(5)	gene.notype; phenotype.pic; allele	(2)	mental; property; state; cognitive
Mind	behavior; state; mental; action	(4)	visual; perception; perceptual; color	(4)	neural.roscience; processing; cortex	(23)	information; dretske; signal
Neuroscience	system; inform.; process; cognitive	(15)	cause.ation; event; intervention	(16)	spirtes; intervention; scheines; pearl	(8)	cause.al; event; variable
Perception	object; experience; perception; color	(7)	model; simulation; datum; measure	(12)	idealize.ation; batterman; approxim.	(12)	model; repres.; system; target
Causation	cause.ation; event; effect	(5)	law; explanation.tory; hempel	(15)	mereolog.; markosian; truthmaking	(11)	explanation.tory; understanding; law
Explanation	model; explanation.tory; account	(19)	entropy; energy; atom; chemical	(17)	nonreductive; kim; physicalist.m	(17)	chemical, chemist.ry; substance
Property	property; world; object; relation	(2)	spacetime; einstein; relativity; clock	(22)	macro.microstate; microcanonical	(22)	measure.ment; scale; quantity
Particles	theory; energy; law; particle	(22)	quantum; particle; measure.; wave	(7)	inertial; spacetime; relativity.istic	(13)	entropy; time; system; state
Quantum	time; state; space; quantum	(17)	kant; newton; galileo; motion	(9)	eigenstate.s; quantum; superpos.	(6)	time; space; theory; relativity
Classics	motion; body; force; newton	(16)	science.tific; philosophy; history	(2)	seventeenth; newton; descartes	(5)	quantum; state; particle; measure.
History	work; time; man; history	(13)	moral; man; emotion; god	(13)	spiritual; conscience; dostoevsky	(9)	newton; motion; galileo; body
Philosophy	world; nature; knowledge; concept	(21)	social.ciety; science; economic	(3)	mankind; lundberg; society; civiliz.	(16)	theory.retical; term; model
Social	science.tific; social; research	(18)	kuhn; popper; laudan; lakatos	(10)	laudan; kuhn; kuhnian; lakatos	(3)	science.tific; theory; research

Discussion

Limitations of the study may concern the corpus used, especially its preprocessing quality and residual noise. Another limitation is our focus on four topic modeling approaches—many others remaining unexplored—and a set of metrics that only cast particular perspectives and all show obvious weaknesses. Nevertheless, the findings revealed significant trade-offs in performance. For example, Top2Vec excels in coherence and diversity but performs poorly in recall and interpretability. LDA and BERTopic perform well in recall but less so in coherence and diversity, favoring broader coverage. CFMf appears to balance these trade-offs effectively.

The study highlighted distinct advantages and drawbacks of the four approaches. Contrary to BOW-based approaches, embedding-based models like Top2Vec and BERTopic rely on text-representation learning: Doc2Vec requires a substantial amount of text to be effective while transformer-based models depend on the very large datasets used for training. Clustering methods also differ significantly. While the BOW-based approaches we tested require choosing the number of clusters beforehand, this can only be done indirectly for embedding-based methods using HDBSAN like Top2Vec and BERTopic, making it more difficult to identify an optimum model based on specific metrics. Also, while LDA performs fuzzy clustering, the other three approaches crisp-cluster documents and interpret clusters

as topics. As a result, handling ambiguity varies among methods. LDA represents documents as probability distributions over topics while Top2Vec and BERTopic rely on HDBSAN for document clustering and outlier detection and deploy specific approaches for outlier reassignment. The angular clustering adaptation implemented with CFMf solved the problem of outlier classes with high document count, and future work will evaluate a more fine-grained outlier reassignment strategy which could also impact small classes.

Extraction of top-words also vary significantly. While LDA simultaneously optimizes probability distributions for topics in documents and for words in topics, the other three approaches extract top-words in a second step after document clustering, for instance through word-topic embeddings distance for Top2Vec, c-TFIDF for BERTopic or Feature Maximization for CFMf. Future work will more systematically explore word ranking and topic profiling using word intrusion tasks.

Conclusion

Overall, the comparative study we conducted shows contrasting results for BOW-based models and embedding-based models. No single approach uniformly outperforms others across all metrics and top-word interpretability, underscoring the need for multiple evaluation perspectives: while Top2Vec reaches highest coherence and diversity scores, it falls behind in terms of recall and qualitative interpretability; BERTopic only slightly outperforms LDA in terms of coherence and diversity, but not recall; as for CFMf with its angular clustering adaptation, it appears to strike a balance between the different metrics, outperforming both LDA and BERTopic in terms of coherence and diversity, though not recall, and generating top-words with high interpretability. These findings show that statistical BOW-based models, far from being obsolete, stand the ground against recent embedding-based methods. They also reveal critical insights into the modularity of topic modeling pipelines.

Acknowledgments

J.-C.L. acknowledges funding from ANRT. F.L. acknowledges funding from Canada Social Sciences and Humanities Research Council (Postdoctoral Fellowships 756-2024-0557, Grant 430-2018-00899). C.M. acknowledges funding from Canada Social Sciences and Humanities Research Council (Grant 430-2018-00899) and Canada Research Chairs (CRC-950-230795).

References

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics* (arXiv:2008.09470). arXiv.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 7819, pp. 160–172). Springer Berlin Heidelberg.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T.

Solorio (Eds.), *Proc. of the 2019 Conf of the North Am. Chap. of the ACL* (pp. 4171–4186). ACL

Fritzke, B. (1994). A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems*, *7*.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv.

Lamirel, J.-C., Dugué, N., & Cuxac, P. (2016). New efficient clustering quality indexes. *2016 International Joint Conference on Neural Networks (IJCNN)*, 3649–3657.

Lamirel, J.-C., Lareau, F., & Malaterre, C. (2024). CFMf topic-model: Comparison with LDA and Top2Vec. *Scientometrics*. *129*, 6387–6405

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, 1188–1196.

Malaterre, C., & Lareau, F. (2022). The early days of contemporary philosophy of science. *Synthese*, *200*(3), 242.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, 44–49.

https://doi.org/10.51408/issi2025_106

Originality in Scientific Titles and Abstracts Can Predict Citation Count

Jack H. Culbert[1], Yoed N. Kenett[2], Philipp Mayr[3]

[1]*jack.culbert@gesis.org*, [3]*philipp.mayr@gesis.org*
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne (Germany)

[2]*yoedk@technion.ac.il*
Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology,
Kiryat Hatechnionm, 3200003, Haifa (Israel)

Abstract

In this research-in-progress paper, we apply a computational measure correlating with originality from creativity science: Divergent Semantic Integration (DSI), to a selection of 99,557 scientific abstracts and titles selected from the Web of Science. We observe statistically significant differences in DSI between subject and field of research, and a slight rise in DSI over time. We model the base 10 logarithm of the citation count after 5 years with DSI and find a statistically significant positive correlation in all fields of research with an adjusted R^2 of 0.13.

Introduction

One aspect of abstracts that likely varies across scientific domains and changes over time is the abstract originality. While some scientific domains have strict norms on abstract formats and content, the increased challenge of a scientific paper getting attention, due to rapid increase in volume of papers with decreased attention span due to information overload (Hołyst, et al., 2024), likely impacts the originality of abstracts. However, the impact of such pressures on abstract writing could have both a facilitative or inhibitory impact on their originality: Abstracts may become more original over time, to compete for a reader's attention more strongly, or they may become less original, to standardize within scientific disciplines and minimize information overload. A possible way to examine these competing hypotheses is by harnessing computational tools that have been recently developed to quantitatively assess the originality of short narratives, particularly an approach called Divergent Semantic Integration.

Cognitive research developed alongside linguistics and natural language processing (NLP) research, as one of the original goals of NLP was to develop a "general theory of human language understanding" which is "linguistically meaningful and cognitively plausible" (Lenci & Padó, 2022). Recent advancements in NLP over the last 10 years have continued to be utilized in modern cognitive research, aided by the rapid development of (large) language models based on deep learning techniques, in particular transformer models.

Divergent Semantic Integration (DSI) (Johnson, et al., 2023) is a computational metric for short textual narratives which was shown to correlate with empirical measures of originality. DSI is computed as the arithmetic mean of cosine distances

between embeddings of sentences from a language model, measuring the overall richness of the language used by the writer in their narrative.

The driving concept is that divergent ideas contained within the text are mapped to distant areas within the embedding space of the model, thereby more diverse concepts are more distant to each other on average than similar or uncreative concepts – resulting in a higher DSI score. Extensive empirical creativity research has highlighted how higher creative individuals exhibit a richer memory structure and are able to more broadly search, expand, and create original ideas (Beaty & Kenett, 2023) (Benedek, Beaty, Schacter, & Kenett, 2023).

This study follows previous research into creativity in science, which has mainly focused on a research paper's metadata, for example: the age of keywords (Azoulay, Zivin, & Manso, 2011), novel or unusual combinations of keywords (Boudreau, Guinan, Lakhani, & Riedl, 2016), referenced articles (Trapido, 2015) or the network centrality between citing and cited papers, (Shibayama & Wang, 2020), the lattermost notably was also found to correlate with citations.

In this study, we compute the DSI of the combined titles and abstracts of papers contained within Clarivate's Web of Science (WoS) from a diverse number of fields and over time, to explore whether there exist trends in originality that correlate with field of research, primary subject classification, bibliometric measures, publication date, or citation count.

Methodology

DSI is computed as the arithmetic mean of the pairwise cosine distance of the embeddings (produced by BERT (Devlin, Chang, Lee, & Kristina, 2019) in layers 6 and 7) of the sentences in a text with each other. The cosine distance is defined as one minus the inner product of the two input vectors. Equivalently this is formulated as, for a text T defined as an ordered list of length $n > 2$ containing sentences s_i, and the embedding vector from the BERT model at layer k represented as $BERT_k(s_i) = \beta_{i,k}$:

$$DSI([s_1, s_2, \dots, s_n]) = \sum_{k_1,k_2 \in \{6,7\}} \sum_{1 \le i < j \le n} \frac{1 - \frac{\beta_{i,k_1} \cdot \beta_{j,k_2}}{\|\beta_{i,k_1}\| \|\beta_{j,k_2}\|}}{4n}$$

We based our code on the codebase provided alongside (Johnson, et al., 2023) and applied this to the combined title and abstract of articles in a snapshot of the WoS. We augmented the original code through refactoring it into a vectorised function that can be applied in a distributed manner against the databases. We computed the DSI of the titles and abstracts, as detailed in the Data section, and then performed a statistical analysis of the DSI against the other variables as detailed in the Results section.

Data

In this study, we obtained the abstracts and bibliometric information from the WoS as of April 2024, provided by the Competence Network for Bibliometrics.[1] From this database we retrieved all subjects with over 10,000 records with classification "Article". Of these we chose subjects which have at least 1000 abstracts with 199-299 spaces, which we assumed correlates to 200-300 words in each abstract. This sampling strategy was chosen to accommodate the long computation time that DSI requires, and to allow for easier analysis of the data.

As mentioned in the discussion of Figure 2 we did not select an equal number of papers per year, which led to an underrepresentation of older papers–for our continuing work we will resample with an even distribution of papers per year and compute the DSI scores for this new dataset.

After this filtering we arrived at a dataset with 1238 candidate subjects, corresponding to approximately 1,238,000 articles, which is ~1.65% of the WoS. After evaluating the scalability of the code, we observed that an abstract of the required length took around 18.2 minutes (after improving the performance of the code), which was mainly attributable to the asymptotic quadratic complexity of computing the pairwise cosine distance over all embeddings generated in the DSI computation.

We took the largest 100 subjects by paper count since 1980 in the WoS and chose a random sample of 1000 articles with 200-300 words in their abstract, these were not balanced to be representative of the number of papers published by year. We appended the abstract to the title (with a full stop in-between) and used this to compute the DSI for each article, ending in a dataset of 100,000 abstracts analysed.

Furthermore, we removed all 443 articles from 2024 from the analysis, as the April edition of the WoS had collected an unrepresentatively small sample for 2024 in the months before the snapshot. This left us with a final dataset of 99,557 records to analyse.

Alongside the DSI scores the following bibliometric information was extracted from the Competence Network for Bibliometrics' version of the WoS: "Primary Subject", "Publication Year", "Citations after 3 Years", "Citations after 5 Years" and "Total Citations". We identified the field of research (field) for each primary subject through correlating with CWTS' NOWT classification[2] and Clarivate's Research Areas,[3] which is visible in Figure 3. Notably in the NOWT classification, the subject Multidisciplinary Sciences was classified into its own field, and we follow this convention, although this leads to a comparatively higher variance for this field due to its smaller size.

Results

The distribution of DSI by fields of research is plotted in Figure 1 (left). We observe a broadly symmetric distribution around the mean for each field, with long tails. We

[1] https://bibliometrie.info

[2] https://www.cwts.nl/pdf/nowt_classification_sc.pdf

[3] https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html

also note a small difference in mean DSI between fields and a similar range to each field. Performing an ANOVA F-test on these categories resulted in statistics $F(5, 99551) = 5936$, $p < 0.01$, $\eta^2 = 0.298$, confirming that the categories have statistically significant differences in means at a 99% confidence level.

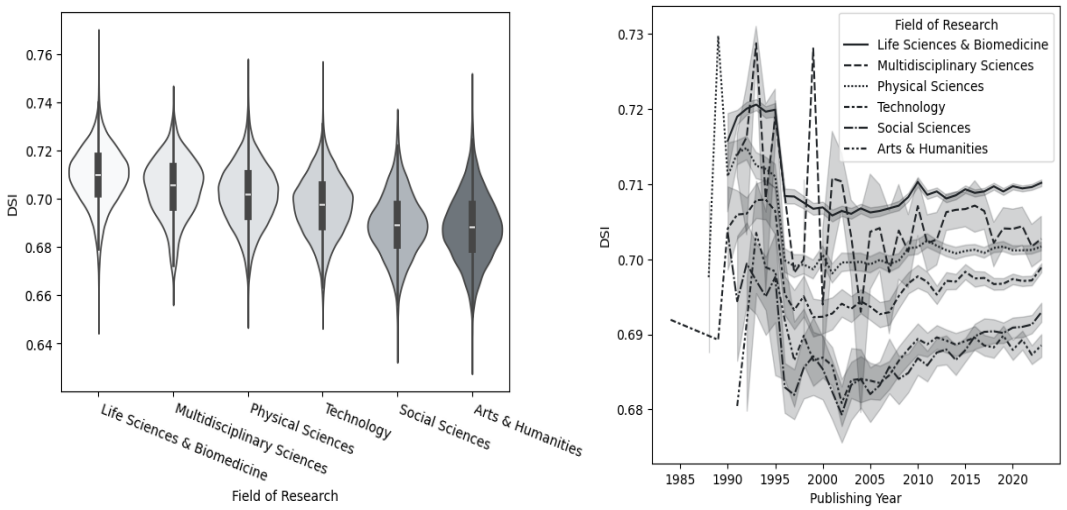


Figure 1. (Left) Violin plots of the DSI for each field, ordered by mean DSI. (Right) Line plots of DSI by publication year with 95% confidence interval, by field.

Observing the progression of DSI per field over time in Figure 1 (right),Figure 2 we see a higher average DSI in the 1990s, which falls and remains stable if not trending slightly positive since 1997 for each field excluding Multidisciplinary Sciences.

Following this observation, we investigated the higher mean and variance of DSI prior to 1997. We found an imbalance of records in our dataset by year–following the well reported global rise in number of papers published by year–which led to an underrepresentation of records the earlier that they were published, due to our random sampling strategy. As mentioned previously, the data will be resampled for following work to correct for this bias.

We modelled citation count using a multilinear model of DSI and field as a categorical variable. We mitigated the bias due to accrual of citations by older papers by correlating the number of citations after 5 years, so for this model we considered only papers published before the end of 2018, to allow for a fair accrual of 5 years of citations before the 2024 sample date. This restriction left us with a dataset of 64,816 records.

As some subjects had a large range in citation count after 5 years, and to better model the large differences in average citation count after 5 years by subject, we took the base 10 logarithm of the citation count after 5 years, (after adding 1 to all citation counts to prevent logarithm errors for papers with no citations).

In Figure 2 we observe a positive correlation between the DSI and base 10 logarithm of the citation count after 5 years for all fields. We performed a statistical analysis of the model: $log_{10}(cit_{5\ years} + 1) \sim DSI + C(Field)$, which was found to be

statistically significant by two-tailed hypothesis test at 99% confidence. The model has a MSE of 0.24, adjusted R^2 of 0.130, Jarque-Bera of 12.918, and a skew and kurtosis of 0.022 and 2.947 respectively. This implies the model explains ~13% of the variation in citation counts. The model may be improved by incorporating publishing year, author count or other bibliometric information, however due to the nature of citation behaviour and the limitations of only analysing titles and abstracts we do not expect a significantly stronger model.

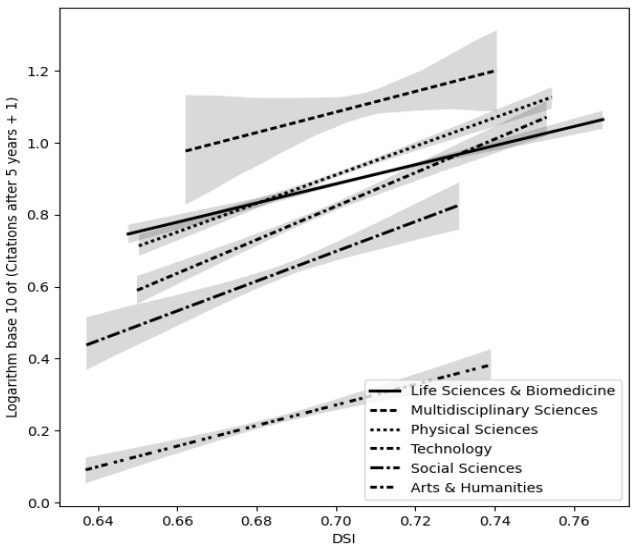


Figure 2. Least Squares Regression for base 10 of the number of citations after 5 years (plus one) as predicted by DSI and field, plotted with 95% confidence interval.

In Figure 3 we break down fields to primary subjects and plot the DSI as a bar chart. We observe broadly similar distributions in DSI across subjects: a unimodal bell-curve with thin, long tails and large overlap of the distribution of DSI between subjects and fields.

In our dataset the five subjects with highest mean DSI in descending order are: Cardiac & Cardiovascular Systems ($\mu = 0.717$, $\sigma = 0.00932$), Ophthalmology ($\mu = 0.715$, $\sigma = 0.0110$), Gastroenterology & Hepatology ($\mu = 0.715$, $\sigma = 0.00986$), Urology & Nephrology ($\mu = 0.714$, $\sigma = 0.00942$) and Obstetrics & Gynecology ($\mu = 0.714$, $\sigma = 0.0111$).

The five subjects with lowest mean DSI in descending order are: Philosophy ($\mu = 0.687$, $\sigma = 0.0137$), Education & Educational Research ($\mu = 0.686$, $\sigma = 0.0122$), Art ($\mu = 0.686$, $\sigma = 0.0132$), Political Science ($\mu = 0.686$, $\sigma = 0.0116$) and History ($\mu = 0.683$, $\sigma = 0.0136$).

Discussion and Conclusion

This large-scale (n = 99,557) ongoing study of the DSI of abstracts and titles in the Web of Science was intended to explore whether this metric, demonstrated in (Johnson, et al., 2023) to be correlated with originality of narratives, also correlates with bibliometric variables.

Our most significant finding so far in this study is our modelling of the logarithm of citation counts after 5 years by DSI and field, which resulted in statistically significant positive correlations which indicate DSI may be a useful computational indicator for future citations (Figure 2).

We observed a statistically significant difference in DSI by field of research, as well as a slight positive trend over time. As there is a large overlapping spread of DSI between fields, this implies that categorising subjects by field may not be the best discriminator for DSI.



Figure 3. Boxplot of DSI scores per subject and field, ordered by mean DSI including outliers and plotted with mean excluding outliers.

We note that subjectively, technologically applied fields appear to have higher DSI than less technologically applied fields. This may be due to the tokenisation and embedding of novel terms creates vectors that do not align with the rest of their field (potentially due to the lack of exposure for the model in training), therefore a next step would be to experiment with a model trained on scientific text such as SciBert (Beltagy, Lo, & Cohan, 2019) for this analysis.

A fundamental limitation of our study is the lack of human-ranked creativity scores for scientific papers, and our assumption that DSI generalises past to scientific ones as a metric of originality. As mentioned previously, our dataset was not balanced in terms of publishing year, which diminishes the strength of our findings in the positive trends of DSI mapped over time.

Furthermore, while DSI was found to generalise across varying language and cultural backgrounds in study 6 of (Johnson, et al., 2023), we have not controlled for English proficiency in this study. Similarly, in study 5 DSI was found to stabilise after 30-50 words up to 200 and was not evaluated at the length we are considering at approximately 200-300 words.

We look to extend this study through analysis of a new collection of data, further analyses of the correlation of DSI with other bibliometric indicators available and computed in the Competence Network for Bibliometrics' version of the Web of Science database to refine our modelling of DSI, as well as experimenting with the embedding model for DSI.

Our results indicate a promising content-based computational method for analysis of scientific papers and potentially a novel link between the creativity sciences and Scientometrics. Computational measures such of these may be of use to the bibliometric community in the analysis of creativity and originality in papers, and perhaps for the wider academic community if this or other originality metrics are incorporated into a search engine as an additional index to re-rank retrieved items.[4]

References

Azoulay, P., Zivin, J. G., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *RAND Journal of Economics*, 527-554.

Beaty, R. E., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in Cognitive Sciences, 27*, 671-683.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: Pretrained Model for Scientific Text. *EMNLP*.

Benedek, M., Beaty, R. E., Schacter, D. L., & Kenett, Y. N. (2023). The role of memory in creative ideation. *Nature Reviews Psychology, 2*, 246-257.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science . *Management Science*, 2765-2783.

Devlin, J., Chang, M.-W., Lee, K., & Kristina, T. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North*

[4] The authors acknowledge funding from the OMINO project (101086321), and Culbert and Mayr additionally acknowledge funding from the OpenBib project (16WIK2301B).

American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota.

Hołyst, J. A., Mayr, P., Thelwall, M., Frommholz, I., Havlin, C., Sela, A., . . . Sienkiewicz, J. (2024). Protect our environment from information overload. *Nature Human Behaviour*.

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., . . . Beaty, R. E. (2023). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods, 55*, 3726--3759.

Lenci, A., & Padó, S. (2022). Editorial: Perspectives for natural language processing between AI, linguistics and cognitive science. *Frontiers in Artificial Intelligence*.

Shibayama, S., & Wang, J. (2020). Measuring originality in science. *Scientometrics*, 409-427.

Trapido, D. (2015). How novelty in knowledge earns recognition: The role of consistent identities. *Research Policy*, 1488-1500.

https://doi.org/10.51408/issi2025_119

Proposal of INDIRECT X Mentions as an Altmetrics Indicator: Dissemination of Research Papers on X via Web News and Blogs

Ai Kishimoto[1], Takayuki Hayashi[2]

[1]*mjs23833@grips.ac.jp, aeki.02081817@gmail.com,* [2]*ta-hayashi@grips.ac.jp*
National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku, Tokyo (Japan)

Abstract

This study reconsiders altmetrics as an indicator for measuring the societal impact of research, specifically focusing on X (formerly Twitter). Current X data is aggregated based on posts containing direct hyperlinks to academic resource, which are first-order citations. In this research, we propose the indicator "INDIRECT X mentions" to capture the dissemination of research through second-order citations via intermediary webpages, such as news article or blogs that describing and linking to the resource. We also compare and validate its effectiveness against existing indicators. Focusing on the AI field, we conducted an analysis using data from one and a half years after publication. The results show that second-order citations allow for measuring the societal impact of papers over a longer period and form a distinct network with almost no overlapping users with the existing networks. Furthermore, the number of intermediary webpages strongly influencing "INDIRECT X mentions" deviates from the overall number of backlinks on the web. A comparison of language proportions also revealed that the proportion of Japanese-language intermediary webpages was significantly lower.

Introduction

The evaluation of academic research has traditionally been based on citation counts and citation-based metrics. These indicators have been widely used not only in academia but also in policymaking. However, concerns have been raised regarding their reliability, time lag before citation counts accumulate and inability to measure the societal impact of research beyond the academic community. Since the 2000s, the widespread use of the Internet has led to increase the need for new evaluation metrics, and J. Priem et al. (2010[a]) proposed the concept of "altmetrics." Today, platforms such as Altmetric.com and PlumX provide real-time indicators beyond citations for assessing research impact.

Among altmetric data sources, X (formerly Twitter) has drawn particular attention due to its data scale and potential as an indicator for measuring the societal impact of research (Wouters, P., et al., 2019). Notably, academic discussions on X have been increasing, and further enhancing its value as a data source (Yu, H., et al., 2019).

J. Priem et al. (2010[b]) stated that the dissemination of academic papers on X involves first-order citations, which hyperlink directly to academic resources, and second-order citations, which hyperlink to intermediary webpages such as news articles or blogs that describe and link to the resources. They found that second-order citations account for up to 48% of mentions. However, existing altmetrics indicators focus exclusively on first-order citations, overlooking second-order citations dynamics.

This study aims to propose a new metric that accurately reflects the dissemination of academic papers on X by incorporating second-order citations. We refer to direct

dissemination through traditional academic platforms as "DIRECT X mentions", while indirect dissemination via intermediary webpages is termed "INDIRECT X mentions". We collect and analyze DIRECT X mentions and INDIRECT X mentions for AI-related papers over 1.5 years post-publication. By examining how papers spread through second-order citations—an aspect largely unexplored in previous studies—we assess the utility of INDIRECT X mentions as a novel metric. To achieve this, we address two research questions.

- RQ1: Can INDIRECT X mentions serve as a new indicator of a paper's social impact?
- RQ2: Are the Altmetrics News and Blogs data reliable?

Method for Collecting INDIRECT X mentions

The existing metric, DIRECT X mentions, counts the total number of posts, including their reposts and quotes, that hyperlink directly to the webpage of an academic resource.

Our new metric, INDIRECT X mentions, counts posts that contain hyperlinks to intermediary webpages that mention the academic resource, as well as their reposts and quotes. This allows INDIRECT X mentions to be aggregated in a manner similar to DIRECT X mentions.

A list of intermediary webpages, such as news articles and blogs that describe and link to academic resources, is aggregated as altmetrics News and Blogs data. By collecting second-order citations that hyperlink to these webpages, we can systematically quantify INDIRECT X mentions.

Data

Several data providers offer altmetrics data, with differences in their collection methods and coverage. Among them, Altmetrics.com is known for its comprehensive coverage of X and webpage data (Ortega, J. L., 2018; Ortega, J. L., 2019; Zahedi, Z., et al., 2018). This study uses Altmetrics.com data to select target papers and refers to the existing X data as "DIRECT X mentions". This study is focusing on AI-related papers as a case study, given the high level of interest in AI research from both researchers and the general public. Papers in the AI field were identified using Altmetrics.com subject codes (4602, 4611) and keywords (AI, Artificial Intelligence, Deep Learning, GPT, LLM, Large Language Model). The selected papers were published between November 1, 2022, and April 30, 2023. For papers on arXiv, the publication date on arXiv was used as the original date. After removing duplicates, the top 100 papers were chosen based on DIRECT X mentions, counted 1.5 years after publication.

For intermediary webpages, News and Blogs data were collected for the 100 papers, duplicates were removed, and shortened URLs were resolved. The data with no URL was excluded, as it originated from non-web news sources such as newspapers. URLs duplicates were identified based on string match. As a result, 4,473 intermediary webpages were selected for analysis.

Next, using NodeXL, a network analysis tool for social media and web data, second-order citations were collected based on the aggregated intermediary webpage list.

Duplicate posts within the same article were excluded, and only posts within the aggregation period were considered. This resulted in 28,926 posts. INDIRECT X mentions were calculated by summing the number of posts, Reposts, and Quotes. Each dataset used in this study has slight differences due to the technical limitations of data collection. Specifically, in INDIRECT X mentions, the reposts and quote counts for second-order citations include posts from private accounts, whereas this is not the case for DIRECT X mentions. However, it was determined that these differences would not significantly affect the study's results.

For the time series analysis, cumulative data for INDIRECT X mentions was created based on post dates, while DIRECT X mentions were collected from the Altmetrics.com timeline. Due to limitations of the original data, the DIRECT X mentions timeline uses the reposted content date, while the INDIRECT X mentions data is added using the date of the original post.

For detailed network analysis, data from 17 selected papers out of the 100 papers were used, including only posts with hyperlinks. The INDIRECT X mentions network dataset contained 17,131 posts. For DIRECT X mentions, data were retrieved from Altmetrics API and X API. Some discrepancies occurred due to privacy settings and account deletions. Of 20,328 tweet_ids, detailed data was available for 20,107 posts. After filtering by date, 5,090 posts remained for analysis. To refine network analysis, DIRECT X mentions were recalculated using the same methodology as INDIRECT X mentions. This recalculated dataset is labelled as "DIRECT X mentions' " to distinguish it from the original data.

Can INDIRECT X mentions serve as a new indicator of a paper's social impact?

Score Distribution and Paper Classification

The minimum value for DIRECT X mentions was 252, while INDIRECT X mentions had a minimum value of 0, with 24 papers receiving no mentions. The first quartile for INDIRECT X mentions was 1, and the median was 90, showing most papers had low mentions. Both datasets had similar distributions, with maximum values of 7,308 (DIRECT X) and 8,625 (INDIRECT X), and interquartile ranges of 615 and 628.5, respectively. The analysis resulted in a Spearman's rank correlation coefficient of 0.53, indicating a moderate positive correlation, though not a strictly proportional relationship. The papers were classified into four quadrants based on the average values of DIRECT X mentions and INDIRECT X mentions (Figure 1): Q1 contained 11 papers, Q2 had 13, Q3 had 64, and Q4 had 12. While Q1 had a limited number of papers, 25 outliers were evenly distributed between Q2 and Q4. Meanwhile, 64 papers were categorized in Q3, indicating an appropriate skew in distribution. The scatter plot revealed a contrasting pattern cantered around $y = x$, supporting the Spearman correlation coefficient of 0.53, which suggests a weak proportional relationship between DIRECT and INDIRECT X mentions. However, the balanced distribution indicates the usefulness of INDIRECT X mentions as an independent metric. Furthermore, the results suggest that a previously overlooked social impact exists at a comparable scale through second-order citations, highlighting 12 papers in Q4 that were not adequately assessed by existing indicators.

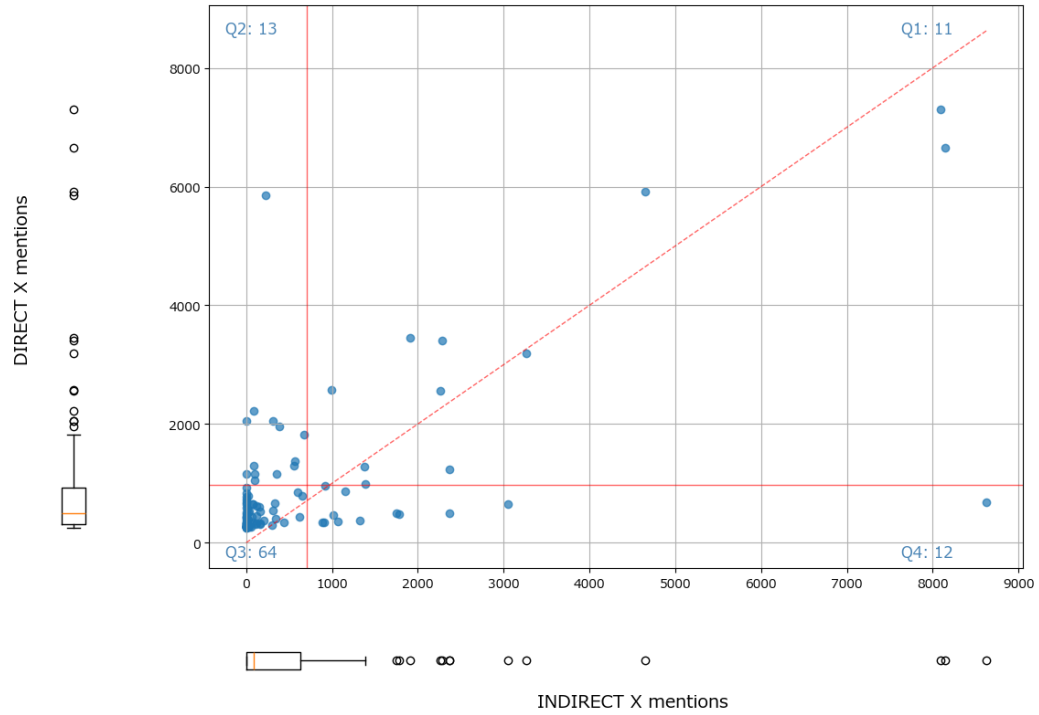


Figure 1. Distribution and Classification of DIRECT/INDIRECT X Mentions.

Time Series Analysis

A time series analysis of DIRECT X mentions and INDIRECT X mentions was conducted to examine their immediacy and long-term influence. One of the features of altmetrics is its rapid response. According to Priem et al.(2010[b]), 15% of first-order citations on X occur on the same day, and 40% occur within a week. In our analysis, 30% of DIRECT X mentions showed a response on the same day, and 90% within a week, while INDIRECT X mentions showed slower responses but still exhibited values comparable to Priem et al.'s findings. A detailed comparison revealed that 77% of papers exhibited a quicker response in DIRECT X mentions than in INDIRECT X mentions, with an average difference of 64 days. INDIRET X mentions may have a slight drawback in terms of immediacy.

To further assess the patterns the progression of DIRECT X mentions and INDIRECT X mentions towards their final scores s was competed. We analysed the data trends for each paper and calculated the median and average number of days to reach thresholds from 50% to 100% based on the final cumulative value. The results (Table 1) indicated that both indicators show a very rapid convergence compared to citation time lags, and INDIRECT X mentions tended to accumulate responses more slowly than DIRECT X mentions. This suggests that INDIRECT X mentions may take more time to gather reactions. Analyzing each paper individually, we found that INDIRECT X mentions tend to continue increasing independently over a long period, regardless of their score magnitude, unlike DIRECT X mentions. This suggests that

even after first-order citations have converged, second-order citations may play a role in maintaining the social impact of the paper.

Table 1. Elapsed days from publication date to threshold attainment date.

	Median		*Mean*	
	DIRECT X	INDIRECT X	DIRECT X	INDIRECT X
50%	7.00 日	41.50 日	42.50 日	103.87 日
60%	10.00 日	52.00 日	46.40 日	113.23 日
70%	12.00 日	76.00 日	59.00 日	127.50 日
80%	31.00 日	89.00 日	78.68 日	152.32 日
90%	66.50 日	127.00 日	125.38 日	188.36 日
100%	502.50 日	378.50 日	460.26 日	297.85 日

Network Analysis

To understand who posted direct or indirect mention, a network analysis based on DIRECT X mentions' (Group1), which recalculated DIRECT X mentions value based on post data obtained from the X API, and INDIRECT X mentions (Group2) was conducted. The 17 papers were randomly selected from each quadrant: 3 papers from Q1 (out of 11), 3 papers from Q2 (out of 13), 3 papers from Q3 (out of 64), and 8 papers from Q4 (out of 12). To examine the network overlap between Group 1 and Group2, Jaccard's Index was used. The results showed that the average Jaccard Index for all papers was 0.012 for Posting User (User who posted) and 0.031 for Referenced User (Referenced user in mentions, replies or quote), indicating a very small overlap. This suggests that the groups who posted or received of the posts of direct hyperlink to academic resource (Group1) and intermediary webpage (Group2) form independent networks.

The characteristics of both networks were analysed by combining data from all the papers. The analysis revealed that, for Posting User, In Group1, it was found that @arxivabs posted 801 posts which accounted for 20% of all posts. Furthermore, to examine users with greater influence, we looked at the X mentions scores for the top 20 users. The 20 users in Group2 accounted for 2.25% of all posts and 24.05% of the total INDIRECT X mentions, while the 19 users in Group1, excluding @arxivabs, accounted for 1.44% of all posts and 45.45% of the total DIRECT X mentions'. When examining Referenced User, it was found that many posts in both Group 1 and Group 2 referred to the same users. However, the most frequently mentioned users in both groups were different (Figure 2): Group 1 included 11 paper authors, 2 institutional accounts, 2 journal accounts, and 5 general users, while Group 2 included 14 news site accounts, 4 paper authors, and 2 corporate accounts. This suggests that Group 1 targets the academic community, whereas Group 2 is more associated with the news business community.

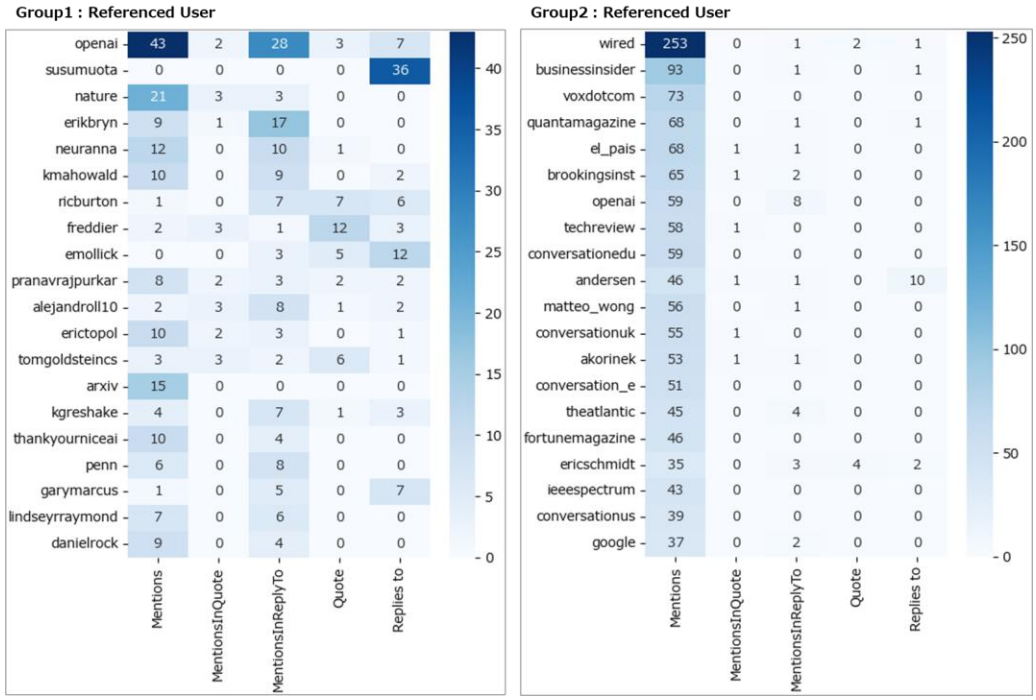


Figure 2. Top 20 Users by number of Referenced Nodes and post count by post types.

Are the Altmetrics News and Blogs data reliable?

Both DIRECT X mentions and INDIRECT X mentions exhibit large fluctuations in values, and no significant correlation was observed. However, INDIRECT X mentions tend to depend on the number of intermediary webpages, so if the number of intermediary webpages is 0, INDIRECT X mentions must be 0. Analysis using Spearman's correlation coefficient showed a strong positive correlation of 0.88 between INDIRECT X mentions and the number of intermediary webpages. In other words, to accurately collect INDIRECT X mentions as proposed in this study, intermediary webpages must be accurate. Currently, News and Blogs data is not automatically tracked for all web sources, but is tracked for those listed in Altmetrics.com's unique site list. While this method ensures the quality of intermediary webpages, it also carries the risk of reducing data collection quality if the site list is incomplete.

Backlink Count Investigation

To verify the reliability of webpages, the number of backlinks for papers with 0 webpages (18 papers) was investigated. Although this investigation includes all backlinks from sources other than news sites and blogs, it revealed that all the papers had numerous backlinks. The number of backlinks ranged from a minimum of 16 to a maximum of 391, with eight papers having more than 100 backlinks. This result strongly suggests that these 18 papers likely have webpages with non-zero values, implying that the current source list or judgment process may be incomplete.

Language Proportion Analysis

Previous research has shown that news and blog data are heavily biased toward English. Although Altmetrics.com has been found to be superior to other data providers in multilingual data collection, we have to examine the language disparities. Due to data limitations, network data from 17 papers was used (DIRECT X mentions'). The analysis revealed that, as in previous studies, English accounted high proportion in all categories. The proportion of Japanese was 1.44% (ranked 6th) in intermediary webpages, 9.40% (ranked 2nd) in INDIRECT X mentions, and 23.22% (ranked 2nd) in DIRECT X mentions. The high proportion of DIRECT X mentions suggests a strong interest in the field of generative AI research within the Japanese-speaking community. In fact, a survey conducted during the data collection period indicated that Japan ranked 3rd in access share to Openai.com, accounting for about 7%, confirming strong interest in generative AI. Furthermore, Japan ranks 2nd in X (formerly Twitter) usage, just behind the United States, suggesting that the data from DIRECT X mentions likely reflects actual language proportions. Given these findings, the low proportion of Japanese in webpages points to a possible inadequacy in Altmetrics.com's data collection methods for Japanese data.

Conclusion

This study attempted to quantify the dissemination of research papers through second-order citations using INDIRECT X mentions and examined whether it serves as a useful new metric, capturing the social impact of papers that traditional DIRECT X mentions may overlook. The results showed INDIRECT X mentions can reveal papers that were significantly disseminated through second-order citations but were not captured by DIRECT X mentions. The time series analysis indicated that, while INDIRECT X mentions—being citations propagated through webpages—lack immediacy, they continue to generate discussions beyond the initial day of citation, reflecting sustained interest in the research.

Network analysis suggested that second-order citations form an independent network with minimal overlap with first-order citations, suggesting that paper dissemination likely occurs in a broader scope, different from the follower-following relationships on X. Furthermore, while first-order citations primarily involve users from the academic community, second-order citations target the news industry community, indicating that second-order citations may be a useful metric for measuring social impact.

In conclusion, INDIRECT X mentions provide a more comprehensive measure of a paper's influence, offering researchers and policymakers a means to evaluate the societal reception of academic work. However, this study also has its limitations. Concerns remain regarding the reliability of webpage data. To ensure complete data coverage, future studies should incorporate web data from multiple platforms. Additionally, further research could explore the correlation between INDIRECT X mentions and academic success, their predictive value, and their potential as an indicator for measuring institutional research promotion efforts.

References

Priem, J., et al. (2010[a]) Altmetrics: A manifesto. http://altmetrics.org/manifesto.

Priem, J. and Costello, K. L. (2010[b]) How and why scholars cite on Twitter. ASIS&T, Volume47, Issue1, November/December, Pages 1-4. https://doi.org/10.1002/meet.14504701201

Wouters, P., Zahedi, Z., Costas, R. (2019). Social Media Metrics for New Research Evaluation. In: Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M. (eds) Springer Handbook of Science and Technology Indicators. Springer Handbooks. Springer, Cham. https://doi.org/10.1007/978-3-030-02511-3_26

Yu, H., et al. 2019. Who posts scientific tweets? An investigation into the productivity, locations, and identities of scientific tweeters. s.l. : Journal of Informetrics, Volume 13, Issue 3 (2019) 841-855. https://doi.org/10.1016/j.joi.2019.08.001

Ortega, J. L. (2018) Reliability and accuracy of altmetric providers: a comparison among Altmetric.com, PlumX and Crossref Event Data. Scientometrics 116, 2123–2138 (2018). https://doi.org/10.1007/s11192-018-2838-z

Ortega, J. L. (2019) The coverage of blogs and news in thethree major altmetric data providers. s.l. : In 17th International conference of the international society for scientometrics and informetrics, Rome, Italy. https://osf.io/pd27h/

Zahedi, Z. and Costas, R. (2018) General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. PLoS ONE,13(5), e0197326. 2018. https://doi.org/10.1371/journal.pone.0197326.

https://doi.org/10.51408/issi2025_144

Regional Patterns of Plagiarism: Evidence from PhD Theses in Russia

Anna Abalkina[1], Alexander Libman[2], Andrey Zayakin[3]

[1]*anna.abalkina@fu-berlin.de*, [2]*alexander.libman@fu-berlin.de*
Freie Universität Berlin (Germany)

[3]*a.zayakin@gmail.com*
The Insider (Latvia)

Abstract

The paper explains the regional patterns of plagiarism in PhD theses (dissertations) in Russia. Using data from more than 108 thousand dissertations dated between 1996 and 2021, which exhibit a significant amount of text similarities with at least 1,000 6-grams, we explore the regional drivers of scientific misconduct and the science characteristics of each region. Applying two sets of linear regressions, we find a strong negative correlation between the share of dissertations with plagiarism and urbanization, the share of ethnic Russians, and the quality of science in the region. Additionally, we identify a strong positive correlation between plagiarism and the level of corruption within a region.

Introduction

The discrepancy in the frequency of academic dishonesty and plagiarism can be attributed to a core-periphery pattern, where scholars from the periphery are more prone to engaging in scientific misconduct. The core-periphery pattern is used to explain the differences in academic community structures and dishonesty practices across countries. Honig and Bedi (2012) conducted an analysis of text similarities among papers presented at the Academy of Management conference and observed that authors affiliated with developing countries were more prone to plagiarism compared to authors from Western countries. This finding was further supported by Citron and Ginsparg (2015), who investigated the reuse of texts available in arXiv and discovered that authors from developing or ex-socialist countries were more inclined to reuse text. Macháček and Srholec (2022) discovered cross-country differences in questionable research practices and identified geographical patterns of publications in predatory journals, which are associated with middle-income countries with relatively large research sectors.

The core-periphery model is used to explain university-specific patterns of academic misconduct, such as scholars from first-tier universities being found to be less susceptible to engaging in academic misconduct. The study by Fanelli et al. (2022) analyzed papers with problematic image duplications and found that the probability of scientific misconduct is higher if the author is affiliated with a low-ranking university. Bagues et al. (2019) also found that questionable practices such as publications in predatory journals are negatively correlated with university ranking. Much less attention is paid to the core-periphery pattern within a country context and regional patterns of scientific misconduct.

We explore the regional variation in scientific misconduct in PhD theses (dissertations) in Russia, a country known for its significant levels of plagiarism in scientific works and dissertations (Guba & Tsivinskaya, 2024), as well as for significant extensive investigations and awareness of such misconduct (Abalkina, 2024).

We look at the regional perspective of misconduct and corruption at the regional level rather than the university level for several reasons. First, corruption exhibits deep regional patterns in Russia. Second, the mobility of faculty members in Russia is relatively low, not only between different regions but also between universities (Sivak & Yudkevich, 2015). This is further compounded by high rates of inbreeding, where 64% of faculty members in Russia have studied at the same university where they are currently employed (Yudkevich et al., 2013). As a result, the organization of the scientific community in Russian universities has a significant local effect. Thus, the prevalence of academic misconduct is primarily attributed to local patterns rather than the transmission of dishonest practices through the mobility of faculty members. Third, since the early 1990s, regional-level analysis has been more appropriate because Russia has experienced four waves of university mergers with the primary goals of establishing universities based on specialized institutes during the transition period, optimizing existing institutions, and establishing federal universities as well as flagship universities (Romanenko & Lisyutkin, 2018). These mergers exhibit a strong geographical pattern and were implemented within the same city or federal region. Regional patterns of scientific dishonesty in Russia remain underexplored in the literature.

Our work contributes to the literature in several ways. We provide an explanation of plagiarism variation at a country level across regions. While previous studies looked at the differences in scientific misconduct prevalence across countries (Honig and Bedi, 2012) or universities (Rudakov et al., 2019), we show that local regional patterns in Russia such as urbanization, corruption, and ethnic structure explain the variation in plagiarism frequency across Russian regions.

Data and methods

The Russian State Library, which by law deposits the texts of dissertations (Ministry of Education and Science of Russia, 2017) contains more than 1,1 million entries of dissertations dated from 1950 to 2021 that were defended in the Soviet Union and Russia. Among these, 508,352 dissertations were defended in Russia between 1996 and 2021, with 486,586 of them having their texts digitized (Russian State Library, n.d.). Dissernet, a network of researchers and journalists dedicated to identifying plagiarism in dissertations and academic papers written in the Russian language, has performed an automated analysis of text reuse among 460 thousand dissertations taken from the Russian State Library (see Figure 1). Dissernet found more than 111 thousand dissertations dated between 1996 and 2021 with the amount of text similarities with at least 1,000 6-grams, i.e. a sequence of six words in the dissertation. To ensure accuracy, commonly used phrases in dissertations, such as “retaining manuscript rights, the work is accomplished in” (“на правах рукописи

работа выполнена в"), were eliminated. Additionally, the reference lists were not taken into account during the analysis.

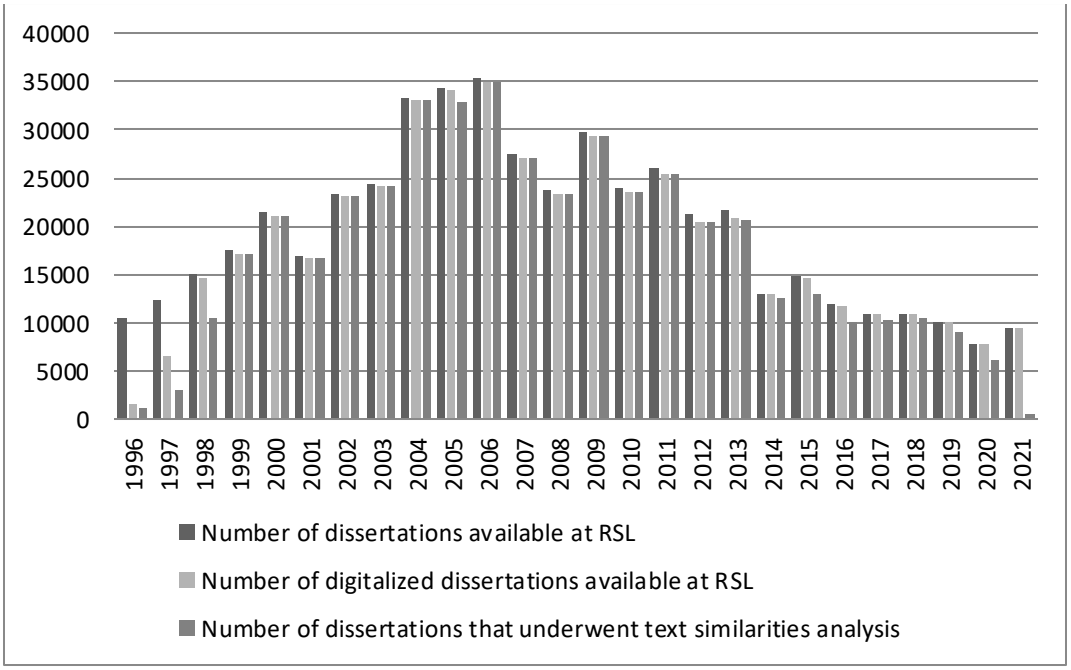


Figure 1. Number of dissertations available at Russian State Library by year.

While the literature suggests that manual checks are necessary to qualify text similarities as plagiarism (Weber-Wulff, 2019), the results of this automated analysis did not undergo manual plagiarism checks to avoid false positives. However, we believe that such identified text similarities are highly likely to be instances of plagiarism. First, Dissernet used a conservative approach by setting a threshold of 1,000 6-grams for text similarities, which exceeds the size used in similar studies, such as Citron and Ginsparg (2015), who used 100 7-grams as a threshold. This conservative approach aims to capture large-scale text similarities and takes into account the relatively high tolerance towards plagiarism in Russia (Rudakov et al., 2019). Second, Dissernet manually checked over 12,500 dissertations (Dissernet, n.d.), establishing that dissertations were the primary source of documented cases of plagiarism in dissertations. Third, duplicates with earlier dates are identified as sources, while the subsequent dissertations are recognized as instances of text reuse. The text similarity analysis was performed in 2022. The region of the dissertation was successfully identified for 108 thousand dissertations, which were included in the subsequent econometric analysis.

In order to analyze regional aspects of misconduct, we perform two sets of linear regressions. The first set of regressions deals with the regional drivers of scientific dishonesty, such as the communist legacy measured as the share of members of the Communist Party in the Soviet Union, urbanization, ethnicity (share of ethnic Russians), and corruption. The second set of regressions takes into consideration the science characteristics of the region, such as the growth of dissertations, the quality

of science measured as a share of publications indexed in the Russian Scientific Citation Index, and the quantity of ideological dissertations during Soviet times. We also control for the regional domestic product and the number of dissertations defended in 2000. The dependent variable is the number of dissertations from 1996 to 2020 with automatically detected text similarities of at least 1,000 6-grams, normalized by the number of dissertations in 2000.

Results

Tables 1 and 2 present results of multiple linear regressions. Regression results reveal that regional patterns are associated with the share of plagiarized dissertations. Specifically, the findings indicate a statistically significant and negative association between urbanization level and plagiarism, suggesting that plagiarism is more common in rural regions. Furthermore, plagiarism is correlated with the ethnic composition of regions, with ethnic regions exhibiting more widespread plagiarism. Additionally, plagiarism, as a form of corruption, is associated with the overall level of corruption in regions, indicating that corruption extends to universities as well.

In some regression specifications, the share of communists in the Soviet Union is also statistically significant in explaining the variation in plagiarism by region. It is known that the share of communists explains variations in corruption and inequality in Russian regions (Libman & Obydenkova, 2021), which is also indirectly related to differences in the spread of plagiarism in Russian regions.

Variation in plagiarism is also associated with scientific patterns. In particular, plagiarism is less prevalent in regions where science is stronger. At the same time, plagiarism is more common in regions where there was a higher growth in dissertation defenses, which apparently indicates a mechanism for the spread of plagiarism. In other words, the increase in the number of dissertations was directly linked to dishonest defenses.

Another aspect of the development of science in Russia is associated with the adaptation of disciplines, especially social sciences, due to the transition to a market economy. Many social sciences were essentially established from scratch in the 1990s. There is evidence that faculty members who previously taught ideological disciplines, such as, for example, the history of the CPSU and dialectical materialism, among others, helped to organize networks of plagiarized dissertations in the newly established social sciences. However, our analysis showed that the share of ideological dissertations defended during the Soviet period turned out to be statistically insignificant in explaining the variation of plagiarism in Russia.

Table 1. Regional patterns of plagiarism variation in Russian regions.

	(1)	(2)	(3)
VARIABLES	**Plagiarism**	**Plagiarism**	**Plagiarism**
Urbanization	-0.714***	-0.527***	-0.341***
	(0.123)	(0.0894)	(0.0753)
Share of communists	1.114**	1.487**	0.534
	(0.487)	(0.576)	(0.409)
Share of ethnic Russians		-16.85***	-7.635***
		(3.812)	(2.611)
Corruption			10.20**
			(4.184)
Constant	62.92***	59.91***	41.99***
	(9.207)	(5.460)	(4.866)
Observations	70	70	63
R-squared	0.508	0.621	0.514

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2. Science pattern of plagiarism variation in Russian regions.

	(1)	(2)	(3)	(4)
VARIABLES	Plagiarism	Plagiarism	Plagiarism	Plagiarism
Urbanization	-0.300***	-0.201**	-0.0926	-0.0872
	(0.0832)	(0.0852)	(0.0815)	(0.0819)
Share of communists	0.501	0.332	0.143	0.150
	(0.399)	(0.404)	(0.381)	(0.383)
Share of ethnic Russians	-7.872***	-9.390***	-9.431***	-9.013***
	(2.467)	(2.290)	(2.175)	(2.345)
Corruption	9.876**	12.17***	13.54***	13.68***
	(4.083)	(4.228)	(3.632)	(3.665)
Log Regional domestic product	-1.603	-2.022	-1.545	-1.465
	(1.999)	(2.088)	(1.596)	(1.609)
Growth of dissertations		0.170**	0.152***	0.148***
		(0.0660)	(0.0474)	(0.0466)
Quality of science (Share of RINC publications)			-0.961***	-0.989***
			(0.181)	(0.187)
Share of ideological dissertations				0.0756
				(0.129)
Constant	59.77**	57.99**	50.15***	48.11**
	(22.45)	(23.46)	(17.69)	(18.41)
Observations	63	63	63	63
R-squared	0.522	0.581	0.683	0.685

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Conclusions

This study explores regional patterns of plagiarism in Russian dissertations. The results of the analysis showed that regional characteristics determined the spread of plagiarism. This research also contributes to the understanding of scientific misconduct through the lens of the core-periphery pattern.

Acknowledgments

Andrey Zayakin thanks Freie Universität Berlin for scholars at risk grant that allowed to design the study.

Competing interests

Andrey Zayakin is a co-founder of Dissernet.

Author contributions

AA – conceptualization, formal analysis, methodology, investigation, visualization, writing – original draft, AL – conceptualization, methodology, supervision, AZ – conceptualization, data curation, formal analysis, methodology, investigation, writing – original draft.

References

Abalkina, A. (2024). Quality and Policies for Academic Integrity: Challenges Faced by Russian Universities. In: Eaton, S.E. (eds) Second Handbook of Academic Integrity. Springer International Handbooks of Education. Springer, Cham. https://doi.org/10.1007/978-3-031-54144-5_174

Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2019). A walk on the wild side: 'Predatory' journals and information asymmetries in scientific evaluations. *Research Policy*, *48*(2), 462-477. https://doi.org/10.1016/j.respol.2018.04.013

Citron, D. T., & Ginsparg, P. (2015). Patterns of text reuse in a scientific corpus. Proceedings of the National Academy of Sciences, 112(1), 25-30. https://doi.org/10.1073/pnas.1415135111

Dissernet (n.d.). Dissernet in exact numbers. Retrieved on 02.07.2023 from https://web.archive.org/web/20230702195941/https://dissernet.org/

Fanelli, D., Schleicher, M., Fang, F. C., Casadevall, A., & Bik, E. M. (2022). Do individual and institutional predictors of misconduct vary by country? Results of a matched-control analysis of problematic image duplications. *PLOS ONE, 17*(3), e0255334. https://doi.org/10.1371/journal.pone.0255334

Guba, K.S., Tsivinskaya, A.O. (2024). Ambiguity in Ethical Standards: Global Versus Local Science in Explaining Academic Plagiarism. *Science and Engineering Ethics*, **30**, 4. https://doi.org/10.1007/s11948-024-00464-6

Honig, B., & Bedi, A. (2012). The fox in the hen house: A critical examination of plagiarism among members of the Academy of Management. *Academy of Management Learning & Education, 11*(1), 101–123. https://doi.org/10.5465/amle.2010.0084

Libman, A., & Obydenkova, A. V. (2021). *Historical legacies of communism: Modern politics, society, and economic development*. Cambridge University Press.

Macháček, V., & Srholec, M. (2022). Predatory publishing in Scopus: Evidence on cross-country differences. *Quantitative Science Studies, 3*(3), 859-887. https://doi.org/10.1162/qss_a_00213

Ministry of Education and Science of Russia (2017). Order of the Ministry of Education and Science of Russia N 1093 "On approval of the Regulations on the Council for the defense of dissertations for the degree of candidate of science, for the degree of doctor of science", November 10, 2017.

Romanenko, K. & Lisyutkin, M. (2018) University Mergers in Russia, Russian Education & Society, 60:1, 58-73, DOI: 10.1080/10609393.2018.1436295

Rudakov, V., Roshina, Ya., Bitokova, L. (2019). Izmeneniya strategiy, motivatsiy i ekonomicheskogo povedeniya studentov i prepodavateley rossiyskikh vuzov (Changes in strategies, motivations and economic behavior of students and teachers of Russian universities). Higher School of Economics.
URL: https://memo.hse.ru/data/2019/03/05/1196154632/2019_inbul_133(1).pdf

Russian State Library (n.d.). General electronic catalogue. Retrieved on 01.07.2023 from https://search.rsl.ru/ru/search

Sivak, E. & Yudkevich, M. (2015). Academic immobility and inbreeding in Russian University sector. In M. Yudkevich, P. G. Altbach, & L. E. Rumbley (Eds.), Academic inbreeding and mobility in higher education. Global perspectives (pp. 130–155). Basingstoke, Palgrave Macmillan.

Weber-Wulff, D. (2019). Plagiarism detectors are a crutch, and a problem. Nature, 567.

Yudkevich, M., Kozmina, Y., Sivak, E., Bain, O., & Davydova, I. (2013). Changing academic profession: Russia Country report. Moscow: Higher School of Economics.

https://doi.org/10.51408/issi2025_094

Research leadership recommendation in research leading-participating multiplex networks based on Wasserstein Distance

Chaocheng He[1], Guiyan Ou[2], Fuzhen Liu[3], Sitong Xiang[4], Ye Zhang[5], Jiang Wu[6]

[1]*he_chaocheng@whu.edu.cn,*
Wuhan University, School of Information Management, Wuhan (China)
Wuhan Unversity Shenzhen Research Institute, Shenzhen, Guangdong, (China)

[2]*Ouguiyan@whu.edu.cn,*[3]*fuzhen.liu@whu.edu.cn,*[4]*2023301043004@whu.edu.cn,*
[5]*2023301043009@whu.edu.cn,*[6]*jiangw@whu.edu.cn*
Wuhan University, School of Information Management, Wuhan (China)

Abstract

Research leadership has long been a central focus in research collaboration. Effective research leadership recommendation is critical for identifying suitable collaborators. However, existing studies predominantly focus on recommending co-author relationships, neglecting the dimension of research leadership. In the context of multiplex networks, existing literature measures interlayer similarity using centrality correlations, which capture only a limited aspect of node importance. To this end, we propose a RMNW model for research leadership recommendation. RMNW constructs a two-layer network: the target layer represents research leadership relationships, while the auxiliary layer captures research participation relationships. The model utilizes Wasserstein Distance to quantify interlayer similarity based on local and global neighborhoods. It integrates information from both layers for link prediction in the target layer, controlled by a tunable parameter λ to balance contributions from each layer. Extensive experiments validate the RMNW model, showing that it significantly outperforms state-of-the-art methods for link prediction in multiplex networks.

Introduction

Research collaboration has become essential due to the growing complexity and nonlinearity of contemporary scientific challenges (Schneider, Sogbanmu et al. 2024). It combines complementary knowledge and expertise from diverse sources to address problems and foster innovation (Gu, Pan et al. 2024). Collaborative efforts typically maintain higher standards of internal quality control compared to single-authored publications (de Frutos-Belizón, García-Carbonell et al. 2024). Collaborator recommendation has received increasing attention across various fields (Liu, Wu et al. 2023, Zhu, Quan et al. 2023).

Research leadership has always been a focal point of research collaboration. Leading authors (first and corresponding authors) play a critical role in securing the academic resources and expertise necessary to initiate and sustain these endeavors (Chinchilla-Rodríguez, Sugimoto et al. 2019). Leading authors primarily offer global,

comprehensive and sustained contributions (Sekara, Deville et al. 2018, Xu, Liu et al. 2024). It is essential for researchers to identify suitable research leaders to initiate and advance collaborative teams for new projects. Similarly, it is crucial for research leaders to select appropriate participating authors (non-first or non-corresponding authors) who can provide local, specialized and staged contributions (He, Wu et al. 2022). Throughout the collaboration process, interactions between leading and participating authors are typically more frequent, closer, and more reciprocal than interactions between two participating authors (He, Liu et al. 2023). Therefore, a two-layer multiplex network, where one layer represents research leadership relationships and the other represents research participation relationships, provides a more effective framework for modeling the complex dynamics inherent in research collaboration.

The recommendation of research leadership (i.e., leading-participating relationships or leading-leading relationships) is critical for effective collaborator identification. However, existing studies face three major limitations. First, most studies focus on recommending relations among all co-authors within collaborations (Liu, Wu et al. 2023), while overlooking the crucial dimension of research leadership relations. Second, although some studies involve research leadership recommendation (He, Liu et al. 2023), they usually model collaboration dynamics using single-layer networks (He, Wu et al. 2021, Cai, Tian et al. 2024), thereby neglecting the interlayer interactions between leadership and participation relationships. Third, existing literature generally measures interlayer similarity by correlating node centralities, such as degree-degree correlation (Zhao, Li et al. 2014) and average similarity of neighbors (ASN) (Najari, Salehi et al. 2019). However, each of these centralities captures different dimensions of node importance, which can introduce bias into the prediction process.

To this end, we propose a novel two-layer Research Leading-Participating Multiplex Network (RLPMN). In the RLPMN, the first layer captures research leadership relations, while the second layer captures research participation relations. Second, we introduce a novel interlayer similarity measure between the target and auxiliary layers, based on the Wasserstein Distance between the local and global neighborhoods of nodes in each layer. In summary, the primary contributions of this study are as follows:

This is the first study to analyze research leadership and collaboration through the lens of a multiplex network.

It is the first to recommend research leadership relations by integrating both intralayer information from the research leadership layer and interlayer information from the research participation layer.

This study introduces a novel approach, namely, adopting Wasserstein Distance to accurately measure the distributional distance of node neighborhoods across layers, providing a precise measure of interlayer similarity.

The proposed framework

Figure 1 illustrates the proposed framework, RMNW (Research leadership recommendation in research leading-participating multiplex networks based on Wasserstein Distance). It is composed of four modules: (1) network construction, (2) interlayer similarity, (3) synthesizer, and (4) recommendation.

Network construction

Let $\boldsymbol{\mathcal{G}} = (G_T, G_A)$ be our proposed Research Leading-Participating Multiplex Network (RLPMN), where $G_T = (V_T, E_T)$ represents the target layer, consisting of research leadership relations (link set E_T), including leading-participating connections and leading-leading connections, which are established between a leading author (either the first author or corresponding author) and a participating author (non-first and non-corresponding), or between two leading authors. And $G_A = (V_A, E_A)$ signifies the auxiliary layer, which comprises research participating relations (link set E_A), including the participating-participating connections between two participating authors (non-first and non-corresponding authors). The common node set $V_T = V_A$ contains N nodes ($|V_T| = N$). Regarding link weights, consistent with the approach of (Zeng, Shen et al. 2017), we assign equal credit to all leading authors (both the first author and corresponding author). In the target layer, the link weight between a leading author i and the $j-th$ author is as follows,

$$W_{ij} = \begin{cases} 2, & if\ j\ is\ also\ a\ leading\ author \\ 1 + \frac{1}{j}, & if\ j\ is\ a\ participating\ author \end{cases} \quad (1)$$

In the auxiliary layer, the weight of the link between i-th author and j-th author is as follows,

$$\mathrm{W}_{ij} = \frac{1}{i} + \frac{1}{j} \quad (2)$$

Figure 1 illustrates the construction of the RLPMN based on two co-authored publications.

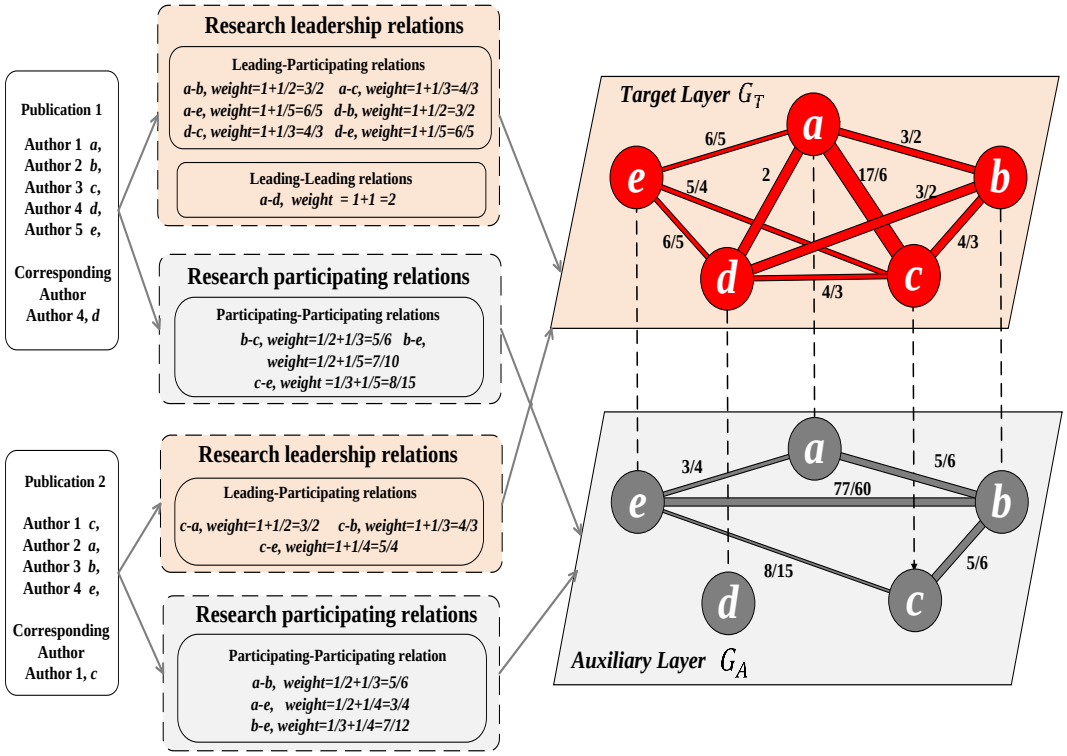


Figure 1. An illustration of Research leading-participating multiplex network (RLPMN) based on two co-authored publications.

Interlayer similarity

To calculate the internal node distance, we embed each node into a low dimensional space. Specifically, as illustrated in Figure 2, the calculation process is divided into three sub-modules: Neighborhood Sampling by 2^{nd} Order Biased Random Walk (NSBRW), Node2Vec embedding, and Interlayer Similarity based on Wasserstein Distance of interlayer neighborhood distribution (ISWN).

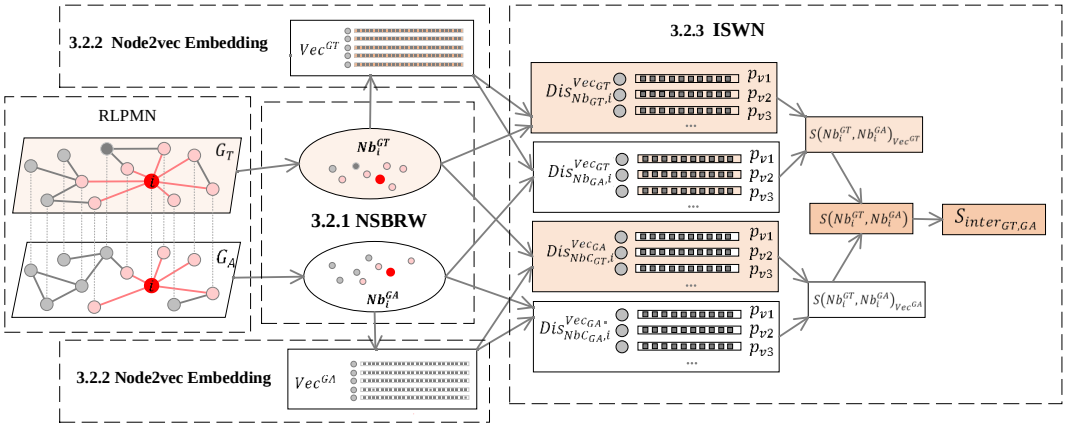


Figure 2. The proposed framework of calculating interlayer similarity.

NSBRW

Following the work of Grover and Leskovec (2016), we employ a 2^{nd} order biased random walk characterized by two parameters p and q. This approach allows for a smooth interpolation between breadth-first sampling (BFS) and depth-first sampling (DFS), enabling the sampling of both immediate and high-order neighborhoods for each node in both the target and auxiliary layers. Figure 3 illustrates the sampling process for the focal node i. From the target layer G_T, we generate Nb_i^{GT}, the neighborhoods of node i, and from the auxiliary layer G_A, we generate the neighborhoods of node i, Nb_i^{GA}.

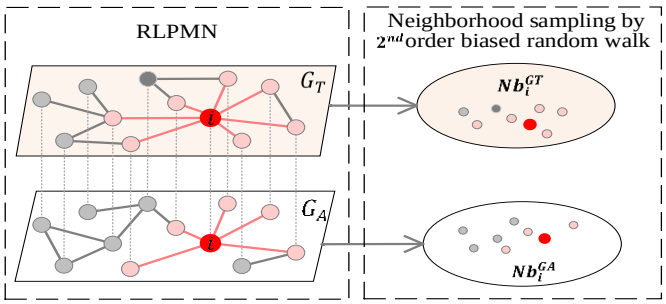


Figure 3. An illustration of neighborhood sampling by 2nd order biased random walk.

Node2vec embedding

With the neighborhoods of each node sampled from both the target and auxiliary layers, we can embed each node into a $d-dimensional$ vector following the node2vec algorithm (Grover and Leskovec 2016), by extending the Skpi-gram architecture to networks. Figure 4 illustrates the Node2vec embedding process for the focal node i. We can embed node i into a d-dimensional vector Vec_i^{GT} based on

the neighborhoods sampled from the target layer G_T, and the a d-dimensional vector Vec_i^{GA} based on the neighborhoods sampled from the auxiliary layer G_A.

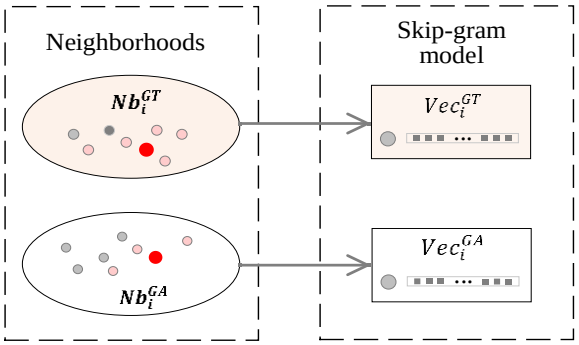


Figure 4. An illustration of Node2vec embedding.

ISWN

In Section 2.2.1, for a focal node i, we have obtained the neighborhoods Nb_i^{GT} sampled from the target layer G_T, and the neighborhoods Nb_i^{GA} sampled from the auxiliary layer G_A. In Section 2.2.2, we have embedded each node into a d-dimensional vector Vec^{GT} based on the target layer G_T, and a d-dimensional vector Vec^{GA} based on the auxiliary layer G_A. Therefore, we can represent neighborhoods Nb_i^{GT} into a d-dimensional space distribution based on the node vector Vec^{GT} as $Dis_{Nb_{GT}}^{Vec^{GT}}$, and into a d-dimensional space distribution based on the node vector Vec^{GA} as $Dis_{Nb_{GT}}^{Vec^{GA}}$. Similarly, we can represent neighborhoods Nb_i^{GA} into a d-dimensional space distribution based on the node vector Vec^{GT} as $Dis_{Nb_{GA}}^{Vec^{GT}}$, and into a d-dimensional space distribution based on the node vector Vec^{GA} as $Dis_{Nb_{GA}}^{Vec^{GA}}$. Consequently, we can obtain the distance between Nb_i^{GT} (the focal node i's neighborhoods from the target layer G_T) and Nb_i^{GA} (i's neighborhoods from the auxiliary layer G_A) via the node embedding vector based on the target layer vector Vec^{GT} by the Wasserstein Distance as follows,

$$Distance\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GT}} = WassersteinD\left(Dis_{Nb_{GT},i}^{Vec^{GT}}, Dis_{Nb_{GA},i}^{Vec^{GT}}\right) \quad (3)$$

The similarity of the focal node i's neighborhoods in the target layer G_T and the auxiliary layer G_A based on the target layer on the target layer vector Vec^{GT} is calculated as follows,

$$S\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GT}} = \frac{1}{1 + Distance\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GT}}} \quad (4)$$

Similarly, we can obtain the distance between Nb_i^{GT} and Nb_i^{GA} via the node embedding vector based on the target layer Vec^{GA} by the Wasserstein Distance as follows,

$$Distance\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GA}} = WassersteinD\left(Dis_{Nb_{GT},i}^{Vec^{GA}}, Dis_{Nb_{GA},i}^{Vec^{GA}}\right) \quad (5)$$

The similarity of the focal node i's neighborhoods in the target layer G_T and the auxiliary layer G_A based on the target layer on the auxiliary layer vector Vec^{GA} is calculated as follows (Segaran 2007),

$$S\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GA}} = \frac{1}{1 + Distance\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GA}}} \quad (6)$$

As shown in Figure 5, the symmetric similarity of the focal node i's neighborhoods in the target layer G_T and the auxiliary layer G_A is calculated as follows,

$$S\left(Nb_i^{GT}, Nb_i^{GA}\right) = \frac{1}{2}\left(S\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GT}} + S\left(Nb_i^{GT}, Nb_i^{GA}\right)_{Vec^{GA}}\right) \quad (7)$$

Finally, the interlayer similarity of the target layer and the auxiliary layer is the mean of similarity of all nodes' neighborhoods in G_T and G_A,

$$S_{inter_{GT,GA}} = \frac{1}{N}\sum_{i=1}^{N} S\left(Nb_i^{GT}, Nb_i^{GA}\right) \quad (8)$$

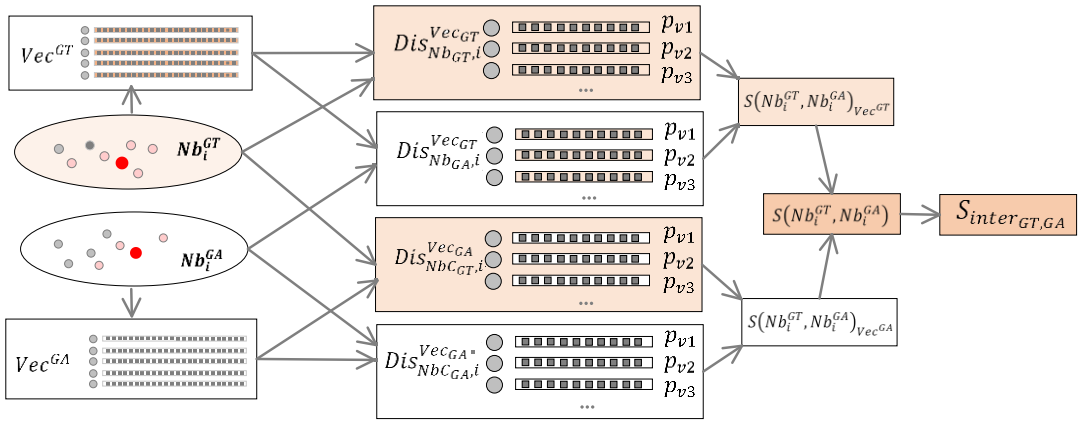


Figure 5. An illustration of Interlayer Similarity based on Wasserstein Distance of interlayer neighborhood distribution.

Synthesizer

For the node pair (u, v) in the target layer G_T, the synthesized index for link possibility is defined as follows (Wu, Ji et al. 2023),

$$P_{u,v} = (1-\lambda) \times P_{u,v}^{G_T} + \lambda \times S_{inter_{GT,GA}} \times P_{u,v}^{G_A} \quad (9)$$

Here, $P_{u,v}^{G_T}$ represents the existence likelihood of the link (u, v) in the target layer G_T based on traditional methods, solely adopting the intralayer information of G_T. Leveraging the node vectors based on each layer in Section 2.2.2, we adopt the cosine

similarity of Vec_u^{GT} (the vector representation of node u based on the target layer G_T) and Vec_v^{GT} (the vector representation of node v based on the target layer G_T) to measure the intralayer similarity of the node u and v in the target layer G_T. Similarly, $P_{u,v}^{G_A}$ denotes the existence likelihood of the link (u, v) in the auxiliary layer G_A, solely based on the intralayer information of G_A. The parameter λ is the tunable variable that determines the weight of information provided by the auxiliary layer G_A for the link prediction in the target layer G_T.

Recommendation

We can conduct research leadership recommendation based on the RLPMN in Section 2.1, interlayer similarity $S_{inter_{GT,GA}}$ in Section 2.2, and the intralayer information $P_{u,v}^{G_T}$, $P_{u,v}^{G_A}$ and synthesizer $P_{u,v}$ in Section 2.3. According to the Equation (9), we can obtain the link possibility of all node pair (u, v), and by sorting, we can obtain the top N recommendations with the highest $P_{u,v}$.

Table 1. The RMNW model.

Algorithm 1: Pseudo-code fo the proposed method: RMNW

Input: Multiplex network $\boldsymbol{\mathcal{G}} = \left(G_T = (V, E_T), G_A = (V, E_A)\right)$, embedding dimensions d, walk length l, number of walks r, window size k, target nodes $Node_t$,

Output: the top-N recommended nodes list for each target node.

Nb^{GT} = GenerateNeighbors(G_T)
Nb^{GA} = GenerateNeighbors(G_A)
Vec^{GT} = Node2VecEmbedding(k, Nb^{GT})
Vec^{GA} = Node2VecEmbedding(k, Nb^{GA})
$S_{inter_{GT,GA}}$ = InterlayerSimilarity(Nb^{GT}, $Nb^{GA}, Vec^{GT}, Vec^{GA}$)
RecommendedList = Recommend(taregtNode)

GenerateNeighbors(G)
  Initialize Nb^G to Empty
  for $w = 0 -> W - 1$ do
    for $v \in V$ do
      Nb_v^{GT} = BiasedRandomWalk(G, v, l)
      Append Nb to Nb^G
  Return Nb^G

InterlayerSimilarity($Nb^{GT}, Nb^{GA}, Vec^{GT}, Vec^{GA}$)
  Initialize $S_{inter_{GT,GA}} = 0$
  for $v \in V$ do
    $Dis_{Nb_{GT},v}^{Vec^{GT}}$ = NeighborDistri(v, Nb_{GT}, Vec^{GT})
    $Dis_{Nb_{GA},v}^{Vec^{GT}}$ = NeighborDistri(v, Nb_{GA}, Vec^{GT})
    $S(Nb_v^{GT}, Nb_v^{GA})_{Vec^{GT}} = \frac{1}{1 + WassersteinD\left(Dis_{Nb_{GT},v}^{Vec^{GT}}, Dis_{Nb_{GA},v}^{Vec^{GT}}\right)}$
    $Dis_{Nb_{GT},v}^{Vec^{GA}}$ = NeighborDistri(v, Nb_{GT}, Vec^{GA})
    $Dis_{Nb_{GA},v}^{Vec^{GA}}$ = NeighborDistri(v, Nb_{GA}, Vec^{GA})
    $S(Nb_v^{GT}, Nb_v^{GA})_{Vec^{GA}} = \frac{1}{1 + WassersteinD\left(Dis_{Nb_{GT},v}^{Vec^{GA}}, Dis_{Nb_{GA},v}^{Vec^{GA}}\right)}$
    $S_{inter_{GT,GA}} += \frac{1}{2}(S(Nb_v^{GT}, Nb_v^{GA})_{Vec^{GA}} + S(Nb_v^{GT}, Nb_v^{GA})_{Vec^{GT}})$
  $Return\ \frac{1}{|V|} \times S_{inter_{GT,GA}}$
Recommend(t)
Initialize $RecommendedList$ to Empty
for $v \in V$ do
  $P_{t,v}^{G_T}$ = cosine (Vec_v^{GT}, Vec_t^{GT})
  $P_{t,v}^{G_A}$ = cosine(Vec_v^{GA}, Vec_t^{GA})
  $P_{t,v} = (1 - \lambda) \times P_{u,v}^{G_T} + \lambda \times S_{inter_{GT,GA}} \times P_{u,v}^{G_A}$
  Append $P_{t,v}$ to $RecommendedList$
$Return$ Sort($RecommendedList$)

Experiments

In this section, we validate the effectiveness of our proposed model RMNW. Specifically, we present the datasets, baseline methods, evaluation metrics, and

implementation details from Section 3.1 to Section 3.4. The research leadership recommendation results are analyzed in Section 3.5.

Dataset

Experiments are conducted on publications the field of "Pharmaceutical Sciences". Publications are retrieved from the Web of Science Core Citation Database, a widely accepted source for studying scientific publications (Yoo, Jung et al. 2024). An advanced search query, "WC = A AND PY = B" is employed, where A is the above Web of Science categories, and B is the publication year "2014-2023". In total, 426708 publications were retrieved. Single-authored publications are excluded. We conduct author name disambiguation following (Sinatra, Wang et al. 2016). Authors with over ten publications are selected (Zhang 2017), resulting in 30286 authors. The dataset is divided into two subset based on publication year: works published before 2022 serve as the training set, while those from 2022 onward form the testing set (Pradhan and Pal 2020, He, Wu et al. 2022).

Evaluation metrics

We employ widely adopted metrics in recommending systems to evaluate the proposed model, namely $F1\ Score$, $nDCG$, and MRR.

(1) $F1\ Score$: $F1\ Score$ is a popular metric to evaluate the performance of a binary classifier. We can divide all the results into four categories: TP (true positive), FP (false positive), TN (true negative), and FN (false negative).

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{12}$$

As shown in Equation 12, the $F1\ Score$ only cares the proportion of true results and false results, ignoring the ranking of the recommended results. Therefore, two more metrics are adopted.

(2) $nDCG$: Let $rel_{R,i}$ represent the graded relevance of the recommended researcher R at position i based on the ground truth data. Discounted cumulative gain (DCG) penalizes highly relevant researchers that appear lower in the recommendation list. The DCG accumulated at a particular rank position p for a recommendation list R is computed as follows:

$$DCG_{R,p} = rel_{R,1} + \sum_{j=2}^{p} \frac{rel_{R,j}}{\log_2(j+1)} \tag{13}$$

The Ideal Discounted Cumulative Gain (IDCG) arranges the results in descending order by relevance and then calculates DCG:

$$IDCG_{I,p} = rel_{I,1} + \sum_{j=2}^{p} \frac{rel_{I,j}}{log_2(j+1)} \tag{14}$$

The normalized DCG ($nDCG$) for a recommendation result R at a specific rank position p is given by the ratio of $DCG_{R,p}$ to $IDCG_{I,p}$ as follows (Järvelin and Kekäläinen 2002),

$$nDCG_{R,p} = \frac{DCG_{R,p}}{IDCG_{I,p}} \tag{15}$$

(3) MRR: Reciprocal Rank (RR) is a metric used in ranking systems to measure how quickly the first relevant item appears in a list of ranked results. Let $rank_i$ be the first relevant result appears at position i in the ranked result list Q. The RR is calculated as

$$RR = \frac{1}{rank_i} \tag{16}$$

If there is no relevant result in Q, the RR is 0. And MRR (Mean Reciprocal Rank) is the mean of RR,

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{17}$$

Baselines

We compare the proposed model with the following state-of-art methods for link prediction in multiplex networks. We also tune the baselines to their best performance for comparison.

(1) LPGRI (Wang, Tang et al. 2023): The model proposes a interlayer similarity, "Global Relevance", using the average Pearson correlation of all node representation vectors from different layers to leverage the information from the auxiliary layer.

(2) MLRW (Nasiri, Berahmand et al. 2021): The model extends the local random walk by leveraging the interlayer and intralayer information, and defines a biased random walk to find the potential link probability in the target layer.

(3) LPIS (Najari, Salehi et al. 2019): The model adopts the AASN-based (asymmetric average similarity of neighbors) correlation $LPIS/AASN$ as interlayer similarity to leverage the information from the auxiliary layer.

(4) MNE (Zhang, Qiu et al. 2018): The model proposes a network embedding approach to jointly represent information of all layers in the multiplex network. But for the task of link prediction, it simply takes the average probability in all

layers as the final probability of a potential link.

(5) RMNW variants: $RMNW_{DDC}$ is a variant of RMNW which uses the degree-degree correlation (Zhao, Li et al. 2014) as the interlayer similarity measure. $RMNW_{ASSN}$ is another variant, which adopts the asymmetric average similarity of neighbors (AASN) to measure the overlap of common neighbors between node pairs across layers, serving as the interlayer similarity (Najari, Salehi et al. 2019). $RMNW_{LO}$ is yet another variant that employs link overlap to quantify the common edges across layers, which serves as an indicator of interlayer similarity (Najari, Salehi et al. 2019).

Implementation details

The parameter settings used for random walk in 2.2.1 NSBRW, and 2.2.2 Node2vec Embedding follow the typical values adopted in (Grover and Leskovec 2016). Specifically, we set vector dimension $d = 128$, and simulate $r = 10$ random walks of fixed length $l = 80$, starting from each node. The window size is set to k=10, and the return parameter $p = 1$ and in-out parameter $q = 1$. As for the tunable parameter λ, we follow the approach outlined in (Jafari, Abdolhosseini-Qomi et al. 2021), and set $\lambda = 0.5$. For the recommendation task, we randomly select 500 authors as the target authors (Pradhan and Pal 2020, He, Wu et al. 2022) and evaluate the performance of our $RMNW$ model alongside other baseline models.

1.1 *Performance comparison*

We compare our proposed $RMNW$ with various baselines. Table 2 reports the $F1$-score and MRR of recommendation performance. Table 3 presents the performance in terms of $nDCG$. Overall, the proposed $RMNW$ model outperforms all.

Regarding the $F1$-score, as shown in Table 2, $RMNW$ achieves the highest $F1$-score for all recommending number N ($F1@5, F1@7$, …, $F1@30$), except for $F1@3$. Notably, $RMNW$ performs the best when $N = 7$ ($F1@7 = 0.1609$), representing an increase of 0.0072 (4.68%) compared to the highest $F1@7$ among the baselines. As N increases, the $F1$-score of $RMNW$ exhibits a gradual downward trend, a pattern also observed among the baselines. Among the baselines, the $MLRW$ generally achieves the highest $F1$-score, followed by $LPGRI$ and $LPIS$. In particular, when $N = 3$, the $F1@3$ of $MLRW$ exceeds that of $RMNW$. Conversely, $RMNW_{DDC}$ and MNE yield the lowest $F1$-score.

Regarding MRR, as shown in Table 2, the $RMNW$ outperforms all baselines. Specifically, it achieves an increase of 0.0383 (7.03%), compared to the highest $F1@7$ among the baselines. $MLRW$ and $LPGRI$ also exhibit high MRR, while $RMNW_{DDC}$ and MNE perform poorly.

In terms of $nDCG$, as shown in Table 3, $RMNW$ achieves the best performance compared to other baselines. It attains the highest $nDCG$ of 0.3329 for the top 3 recommendation. Among the baseline models, $MLRW$ achieves the highest $nDCG$, followed by $LPGRI$ and $LPIS$. Consistent with the performance in terms of $F1$-score and MRR, MNE and $RMNW_{DDC}$ yield the lowest $nDCG$.

Table 2. $F1$ and MRR of $RMNW$ and baseline methods.

Method	$F1@3$	$F1@5$	$F1@7$	$F1@10$	$F1@15$	$F1@20$	$F1@25$	$F1@30$	MRR
$LPGRI$	0.0929	0.1472	0.1526	0.1472	0.1386	0.1187	0.1032	0.0759	0.5349
$MLRW$	**0.0962**	0.1497	0.1537	0.1488	0.1403	0.1221	0.1058	0.0793	0.5448
$LPIS$	0.0910	0.1419	0.1463	0.1417	0.1343	0.1152	0.1009	0.0746	0.5102
MNE	0.0878	0.1307	0.1332	0.1297	0.1225	0.1032	0.0931	0.0695	0.4445
$RMNW$	0.0933	**0.1584**	**0.1609**	**0.1557**	**0.1468**	**0.1246**	**0.1098**	**0.0812**	**0.5831**
$RMNW_{DDC}$	0.0859	0.1292	0.1311	0.1284	0.1217	0.1028	0.0914	0.0660	0.4401
$RMNW_{ASSN}$	0.0896	0.1365	0.1414	0.1372	0.1301	0.1112	0.0969	0.0712	0.4865
$RMNW_{LO}$	0.0881	0.1334	0.1385	0.1329	0.1270	0.1083	0.0936	0.0697	0.4692

Table 3. $nDCG$ of $RMNW$ and baseline methods.

Method	$nDCG@3$	$nDCG@5$	$nDCG@7$	$nDCG@10$	$nDCG@15$	$nDCG@20$	$nDCG@25$	$nDCG@30$
$LPGRI$	0.2889	0.2832	0.2787	0.2718	0.2684	0.2653	0.2607	0.2617
$MLRW$	0.3068	0.2999	0.2953	0.2894	0.2830	0.2762	0.2715	0.2664
$LPIS$	0.2798	0.2724	0.2679	0.2634	0.2588	0.2533	0.2488	0.2465
MNE	0.2552	0.2526	0.2506	0.2477	0.2452	0.2421	0.2365	0.2329
$RMNW$	**0.3329**	**0.3266**	**0.3228**	**0.3183**	**0.3132**	**0.3089**	**0.3037**	**0.3001**
$RMNW_{DDC}$	0.2546	0.2493	0.2455	0.2425	0.2394	0.2355	0.2303	0.2261
$RMNW_{ASSN}$	0.2713	0.2643	0.2610	0.2564	0.2517	0.2486	0.2434	0.2409
$RMNW_{LO}$	0.2606	0.2568	0.2540	0.2519	0.2479	0.2458	0.2420	0.2396

Sensitivity analysis

In this section, we implement the sensitivity analysis of the parameter λ on the performance of RMNW. We adopt $F1@7$, MRR, and $nDCG@7$ as evaluation metrics with $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$. Figure 6-8 report the results in terms of $F1@7$, MRR, and $nDCG@7$, respectively. Regarding $F1@7$, as shown in Figure 6, $\lambda \in \{0.3, 0.4, 0.5, 0.6\}$ leads to good performance. Conversely, excessively small or large λ would degrade the performance in all three research fields. For MRR, as shown in Figure 7, $\lambda \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ yields better recommending performance. On the one hand, a small λ fails to capture sufficient information from the auxiliary layer. On the other hand, a large λ can also overlook critical information from the target layer. In terms of $nDCG@7$, as shown in Figure 8, $\lambda \in \{0.3, 0.4, 0.5, 0.6\}$ can achieve better recommending performance. Similar to the other metrics, both small and large λ can degrade the performance.

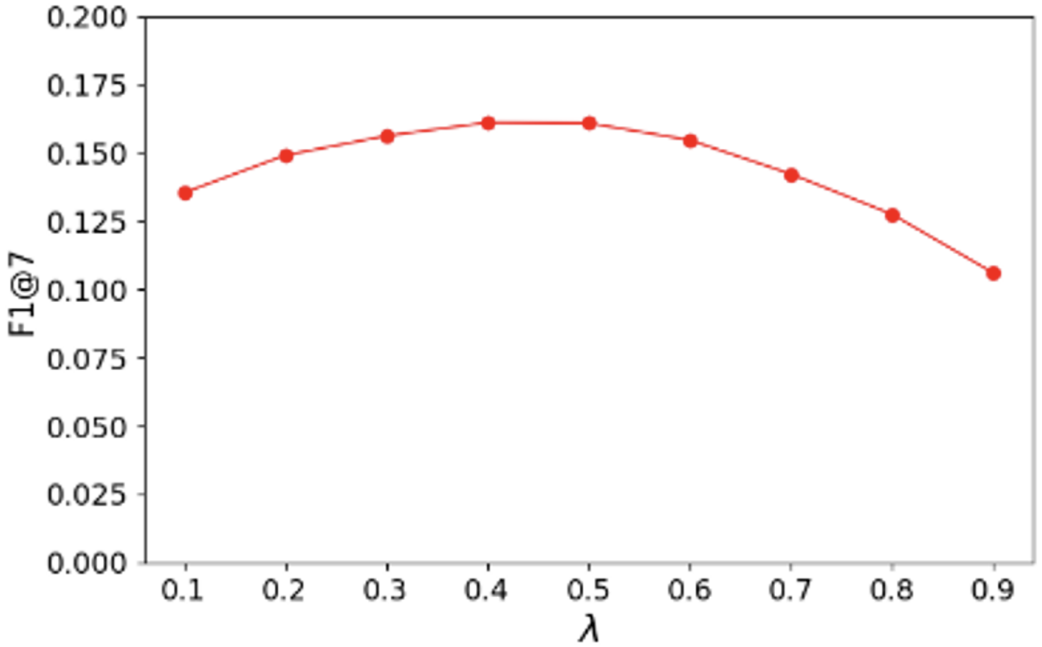


Figure 6. Sensitivity of the parameter λ on the $F1@7$ of RMNW.

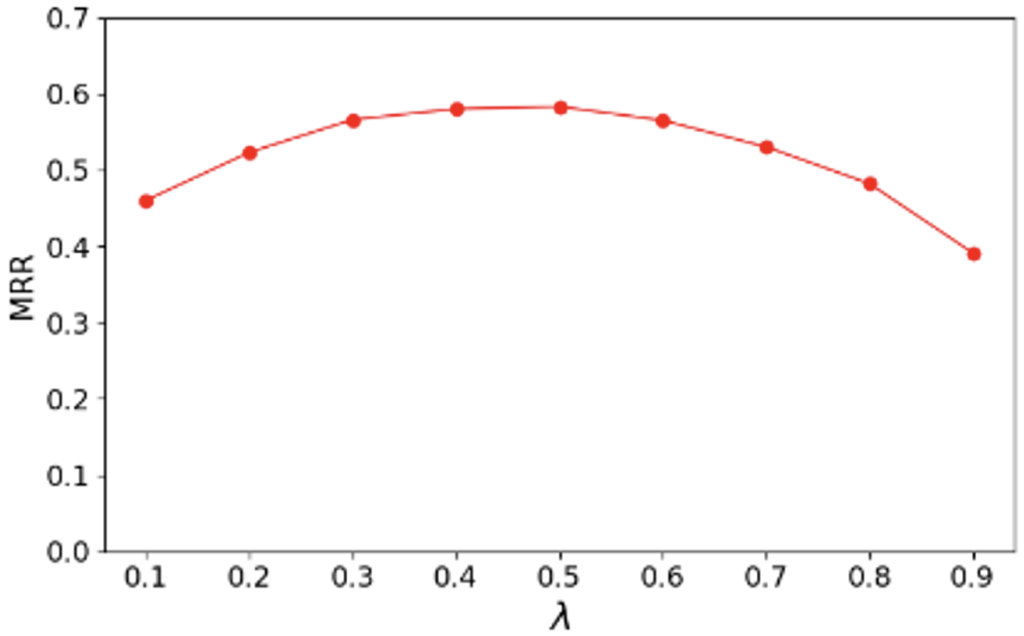


Figure 7. Sensitivity of the parameter λ on the MRR of RMNW.

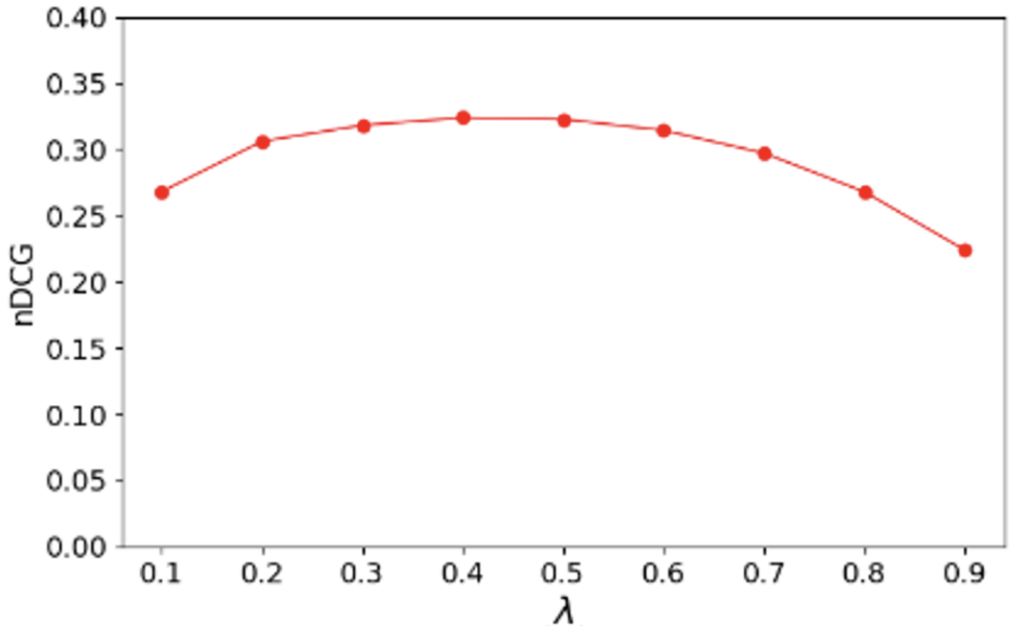


Figure 8. Sensitivity of the parameter λ on the $nDCG@7$ of RMNW.

Ablation studies

We conduct ablation studies to understand the impact of information from the target layer G_T and auxiliary layer G_A on the recommendation performance. The results are presented in Table 4. Model 1 utilizes the information entirely from the target layer G_T ($\lambda = 0$). Model 2 integrates the information from both target layer G_T and auxiliary layer G_A ($\lambda = 0.5$). Model 3 utilizes the information entirely from the auxiliary layer G_A ($\lambda = 1$). From Table 4, we observe that Model 3 consistently yields the lowest $F1@7$, MRR, and $nDCG@7$ across all fields. While Model 2 achieves the best performance. These findings indicate that integrating the information from both layers significantly improves recommendation performance in the target layer G_T.

Table 4. Research leadership recommendation with information from different layers.

Model	λ	Information from the target layer G_T	Information from the auxiliary layer G_A	$F1@7$	MRR	$nDCG@7$
1	0	√	×	0.1182	0.4070	0.2210
2	1	×	√	0.0791	0.2933	0.1513
3	0.5	√	√	0.1609	0.5831	0.3228

Discussion

Our proposed RMNW model significantly improves overall recommendation performance across various metrics, including $F1, MRR, and\ nDCG$. As shown in Tables 2, the $RMNW$ model outperforms baseline models in all evaluation metrics.

Second, integrating information from the research participation layer (auxiliary layer) notably enhances research leadership recommendation performance in the research leading-participating multiplex networks. For example, as highlighted in Tables 3 of the Ablation studies Section, combining information from both the target layer and auxiliary layer ($\lambda = 0.5$) yields improvements of 36.1%, 43.2%, and 46.1% in $F1, MRR, and\ nDCG$, respectively, compared to solely using information from the target layer. Furthermore, even using the degree-degree correlation as interlayer similarity metric, the $RMNW_{DDC}$ achieves better performance than Model 1 ($\lambda = 0$, single-layer network link prediction).

Third, the Wasserstsin Distance effectively captures the interlayer similarity between the target and auxiliary layer. As detailed in Table 2-3, the $RMNW$ consistently achieves the highest $F1, MRR\ and\ nDCG$ compared to other variants such as $RMNW_{DDC}$ (degree-degree correlation), $RMNW_{ASSN}$ (overlap of common neighbors), and $RMNW_{LO}$ (overlap of the common edges).

We also implement the sensitivity analysis of the parameter λ on the recommendation performance ($F1@7, MRR\ and\ nDCG@7$). As depicted in Figure 6, generally, $\lambda \in \{0.3, 0.4, 0.5, 0.6\}$ leads to good performance. And these results confirm the robustness and effectiveness of the proposed RMNW model.

Conclusion and future work

Our work focuses on leveraging research participation relationships to enhance research leadership recommendations. We propose the RMNW model, which consists of four interconnected modules: (1) Network construction. This module distinguishes between research leadership and research participation relationships. It constructs a two-layer network, where the target layer represents research leadership relationships, and the auxiliary layer captures research participation relationships. (2) Interlayer similarity. This module employs the Wasserstein Distance to measure the interlayer similarity based on local and global neighborhoods of nodes in each layer. (3) Synthesizer. This module integrates information from both the target and auxiliary layers for link prediction in the target layer, controlled by a tunable parameter λ. (4) Recommendation. This module identifies potential research leadership partners by ranking the link probabilities of all node pairs and generating the top N recommendations. Extensive experimental results demonstrate that the

RMNW model significantly outperforms state-of-the-art multiplex network link prediction models. Sensitivity analysis further confirms the robustness and effectiveness of the proposed model. Moreover, ablation studies reveal that incorporating information from the research participation layer (auxiliary layer) substantially enhances the performance of research leadership recommendations.

This study has certain limitations that warrant further investigation. In constructing the research leading-participating multiplex network, we do not account for the temporal attribute of the collaboration relationships. However, recent collaborations are more likely to influence future collaboration and should ideally be given greater weight. In subsequent research, we aim to incorporate temporal attributes of collaboration relationships to further improve recommendation performance.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (72204189), Guangdong Basic and Applied Basic Research Foundation (2022A1515110972) and Digital Intelligence Humanities Fundation of Wuhan University (2024SZWK023)

References

Cai, R., W. Tian, R. Luo and Z. Hu (2024). "The generation mechanism of research leadership in international collaboration based on GERGM: a case from the field of artificial intelligence." Scientometrics: 1-19.

Chinchilla-Rodríguez, Z., C. R. Sugimoto and V. Larivière (2019). "Follow the leader: On the relationship between leadership and scholarly impact in international collaborations." Plos one 14(6): e0218309.

de Frutos-Belizón, J., N. García-Carbonell, F. Guerrero-Alba and G. Sánchez-Gardey (2024). "An empirical analysis of individual and collective determinants of international research collaboration." Scientometrics 129(5): 2749-2770.

Grover, A. and J. Leskovec (2016). node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining.

Gu, J., X. Pan, S. Zhang and J. Chen (2024). "International mobility matters: Research collaboration and scientific productivity." Journal of Informetrics 18(2): 101522.

He, C., F. Liu, K. Dong, J. Wu and Q. Zhang (2023). "Research on the formation mechanism of research leadership relations: An exponential random graph model analysis approach." Journal of Informetrics 17(2): 101401.

He, C., J. Wu and Q. Zhang (2021). "Characterizing research leadership on

geographically weighted collaboration network." Scientometrics 126: 4005-4037.
He, C., J. Wu and Q. Zhang (2022). "Proximity-aware research leadership recommendation in research collaboration via deep neural networks." Journal of the Association for Information Science and Technology 73(1): 70-89.
Jafari, S. H., A. M. Abdolhosseini-Qomi, M. Asadpour, M. Rahgozar and N. Yazdani (2021). "An information theoretic approach to link prediction in multiplex networks." Scientific Reports 11(1): 13242.
Järvelin, K. and J. Kekäläinen (2002). "Cumulated gain-based evaluation of IR techniques." ACM Transactions on Information Systems (TOIS) 20(4): 422-446.
Liu, X., K. Wu, B. Liu and R. Qian (2023). "HNERec: Scientific collaborator recommendation model based on heterogeneous network embedding." Information Processing & Management 60(2): 103253.
Najari, S., M. Salehi, V. Ranjbar and M. Jalili (2019). "Link prediction in multiplex networks based on interlayer similarity." Physica A: Statistical Mechanics and its Applications 536: 120978.
Nasiri, E., K. Berahmand and Y. Li (2021). "A new link prediction in multiplex networks using topologically biased random walks." Chaos, Solitons & Fractals 151: 111230.
Pradhan, T. and S. Pal (2020). "A multi-level fusion based decision support system for academic collaborator recommendation." Knowledge-Based Systems 197: 105784.
Schneider, M. D., T. O. Sogbanmu, H. Rubin, A. Bortolus, E. E. Chukwu, R. Heesen, C. L. Hewitt, R. Kaufer, H. Metzen and V. Mitova (2024). "Science–policy research collaborations need philosophers." Nature human behaviour: 1-2.
Segaran, T. (2007). Programming collective intelligence: building smart web 2.0 applications, " O'Reilly Media, Inc.".
Sekara, V., P. Deville, S. E. Ahnert, A.-L. Barabási, R. Sinatra and S. Lehmann (2018). "The chaperone effect in scientific publishing." Proceedings of the National Academy of Sciences 115(50): 12603-12607.
Sinatra, R., D. Wang, P. Deville, C. Song and A.-L. Barabási (2016). "Quantifying the evolution of individual scientific impact." Science 354(6312): aaf5239.
Wang, C., F. Tang and X. Zhao (2023). "LPGRI: a global relevance-based link prediction approach for multiplex networks." Mathematics 11(14): 3256.
Wu, Y., Y. Ji and F. Gu (2023). "Identifying firm-specific technology opportunities in a supply chain: Link prediction analysis in multilayer networks." Expert Systems with Applications 213: 119053.
Xu, H., M. Liu, Y. Bu, S. Sun, Y. Zhang, C. Zhang, D. E. Acuna, S. Gray, E. Meyer and Y. Ding (2024). "The impact of heterogeneous shared leadership in scientific teams." Information Processing & Management 61(1): 103542.

Yoo, H. S., Y. L. Jung, J. Y. Lee and C. Lee (2024). "The interaction of inter-organizational diversity and team size, and the scientific impact of papers." Information Processing & Management 61(6): 103851.

Zeng, A., Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang and H. E. Stanley (2017). "The science of science: From the perspective of complex systems." Physics Reports 714: 1-73.

Zhang, H., L. Qiu, L. Yi and Y. Song (2018). Scalable multiplex network embedding. IJCAI.

Zhang, J. (2017). "Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction." Information Processing & Management 53(1): 42-51.

Zhao, D., L. Li, H. Peng, Q. Luo and Y. Yang (2014). "Multiple routes transmitted epidemics on multiplex networks." Physics Letters A 378(10): 770-776.

Zhu, Y., L. Quan, P. Y. Chen, M. C. Kim and C. Che (2023). "Predicting coauthorship using bibliographic network embedding." Journal of the Association for Information Science and Technology 74(4): 388-401.

https://doi.org/10.51408/issi2025_101

Research on the Measurement Method of Disciplinary Diversity Based on Lexical Semantic Analysis

Guo Chen[1], Yifan Yang[2]

[1] *delphi1987@qq.com*, [2]*1005104368@qq.com*
NJUST Nanjing University of Science and Technology, No. 200 Xiao Ling Wei, Nanjing, Jiangsu (China)

Abstract

Existing methods for measuring disciplinary diversity mainly focus on literature (such as citations) as the unit of analysis. This paper proposes a new approach to measuring disciplinary diversity at a fine-grained level based on lexical semantics. Taking articles from the OpenAlex dataset between 2014 and 2023 as an example, the breadth of concept distribution is calculated within the semantic space of given disciplinary vocabulary to measure disciplinary richness; the external word frequency ratio and similarity of high-frequency disciplinary vocabulary are integrated to calculate the concept overflow degree, thereby measuring the degree of disciplinary intersection. Based on this, a two-dimensional matrix is constructed to locate types of disciplinary diversity and further analyze the temporal trends and causes of diversity in various disciplines. According to disciplinary richness and intersection, 19 first-level disciplines are categorized into four major types: Diverse Integration, Deep Specialization, Broad Interaction, and Single Cohesion, and the classification results are analyzed. Additionally, the trends and causes of changes in richness and intersection at both macro and micro levels are analyzed for each discipline. This study proposes a more fine-grained disciplinary diversity measurement method at the lexical semantic level, providing a new and broader perspective for the study of disciplinary diversity.

Introduction

Traditional methods for measuring disciplinary diversity primarily use literature as the basic unit of analysis, and there is still room for refinement from a fundamental granularity perspective. Words are the fundamental units of knowledge expression, and using their semantics can provide a deeper understanding of the structure and differences in human knowledge content. In psychology, researchers have begun to use word semantics to conduct cognitive experiments. For example, Olson et al. (2021) used cosine distance to calculate the pairwise semantic distances between 10 nouns to measure human divergent thinking, finding it more effective than traditional alternative uses tasks and bridging associative gap tasks. Their findings, published in Nature, have garnered widespread attention. This has inspired many scholars to conduct related work. Hubert et al. (2024) also used the same method to measure the degree of human thinking divergence. Similarly, in addition to measuring human creativity and divergent thinking, word semantics can also be used to measure differences in knowledge. Hur (2024) introduced semantic heterogeneity based on word embedding techniques in content analysis when calculating the diversity of patent entities, representing diversity through the semantic distance between patent entities. Lix et al. (2022) used word semantics to calculate the diversity of team discourse, a concept of fine-grained knowledge participation that is difficult to track

with previous text analysis methods. Thus, it appears possible to use word semantics to reveal disciplinary diversity, but similar research has not yet been conducted.

Words are the most basic units for representing semantics, and in the process of inheriting, communicating, and diffusing scientific knowledge, the finest granularity unit is the conceptual knowledge described by words. Therefore, measuring disciplinary diversity from the perspective of the aggregation and intersection of word semantics is both a natural and inevitable requirement. Based on this, this paper takes word semantics as the starting point, utilizes word semantic representation and deep learning techniques, and analyzes disciplinary diversity from a finer-grained lexical level. It comprehensively considers word frequency and semantic relationships between words, quantifying disciplinary diversity from two dimensions: disciplinary richness and disciplinary intersection. The two dimensions are combined to classify types of disciplinary diversity. In the experimental section, a semantic space for 19 first-level disciplines is constructed using the open-source OpenAlex data, and the proposed method is applied to classify diversity types and analyze time series trends. The empirical results demonstrate that this method can effectively analyze the development characteristics and changes in the degree of intersection of different disciplines, providing a novel perspective and approach for disciplinary evaluation and prediction research.

Data and methods

We used the paper data from OpenAlex between 2014 and 2023 as the experimental subjects, obtaining a total of 72 million records. First, we classified the major disciplines based on the fos (field of study) field in the paper data. If a paper's fos field contains multiple disciplines, it is included in multiple major disciplines. According to the Microsoft discipline classification, there are 19 first-level disciplines. Each discipline is divided into subsets based on the year, resulting in a total of 190 subsets.

The text content undergoes stemming and keyword matching, and the Word2Vec model is trained using incremental learning. The frequency of a word's appearance in different disciplines is used to determine whether it is a discipline-specific term. In this paper, words that appear fewer than nine times are designated as discipline-specific terms for use in subsequent metric calculations.

Currently, the measurement of disciplinary diversity is typically focused at the literature level, resulting in a relatively coarse research granularity that fails to capture subtle semantic changes. However, more fine-grained lexical semantic analysis has been successfully applied to measure the degree of individual divergent thinking and team diversity, indicating that lexical semantic analysis has a solid foundation for representing diversity. Lexical items are the most basic units for representing disciplinary knowledge, and semantic changes can directly explain the development and evolution of disciplinary knowledge. The broader the distribution of vocabulary in a semantic space within a discipline, the richer the disciplinary knowledge is. Therefore, this study employs lexical semantics to measure disciplinary diversity from two key dimensions: the richness within disciplines and the intersection between disciplines. From a semantic perspective, disciplinary

richness can be represented by the average distance between high-frequency words; the greater the average distance, the higher the internal richness of the discipline. Intersection can be represented by the degree of overlap in semantic space; the greater the overlap, the higher the external intersection between disciplines.

Measurement of disciplinary richness

We can measure the average distance between each word and other words to obtain the average distance of elements within the semantic space (or the distance between each word and the document centroid), which can be used to measure the conceptual breadth within that semantic space. Let N be the total number of high-frequency words, v_i and v_j be the word vectors obtained through word embedding, and f_i and $f_j$be the word frequencies.

$$2 \times \frac{\sum_{k=1}^{N} \sum_{i \neq j} \frac{distance\,(v_i \times f_i, v_j \times f_j)}{f_i + f_j}}{N(N-1)}$$

Measurement of interdisciplinary

Combining the word similarity calculation metric and the True Diversity metric (Zhang, L et al., 2016), we have proposed a disciplinary intersection metric based on high-frequency word calculations.For the calculation of disciplinary intersection, let n be the total number of fields, and N be the total number of high-frequency words,. For two different fields i and j, w_{ki} and w_{kj}represent the same words appearing in different fields. p_{ki} and $p_{kj}$are the proportions of the words w_{ki} and w_{kj}in the high-frequency word set N of fields i and j, respectively.

$$\frac{\sum_{k=1}^{N} \sum_{j \neq i}^{n} \mathrm{Cos}\left(w_{ki}, w_{kj}\right) p_{ki}\, p_{kj}}{n-1}$$

Building on the aforementioned approach, lexical semantic calculations can be used to determine the richness within disciplines and the intersection between disciplines. These two metrics can be employed for both two-dimensional matrix analysis and time-series trend analysis. First, a two-dimensional matrix can be used to categorize disciplines into four types, and the possible reasons for these classifications can be analyzed. On the other hand, time-series trend analysis can be conducted to examine the changes in disciplinary richness and intersection over time, and further analysis can be performed from a lexical perspective to understand the reasons for these changes.

Results

Analysis of Lexical Semantic Dimensionality Reduction Visualization Results

To more intuitively observe the distribution of word vectors, this paper employs the UMAP dimensionality reduction algorithm to visualize the distribution of vocabulary from various disciplines, as shown in Figure 1 below.

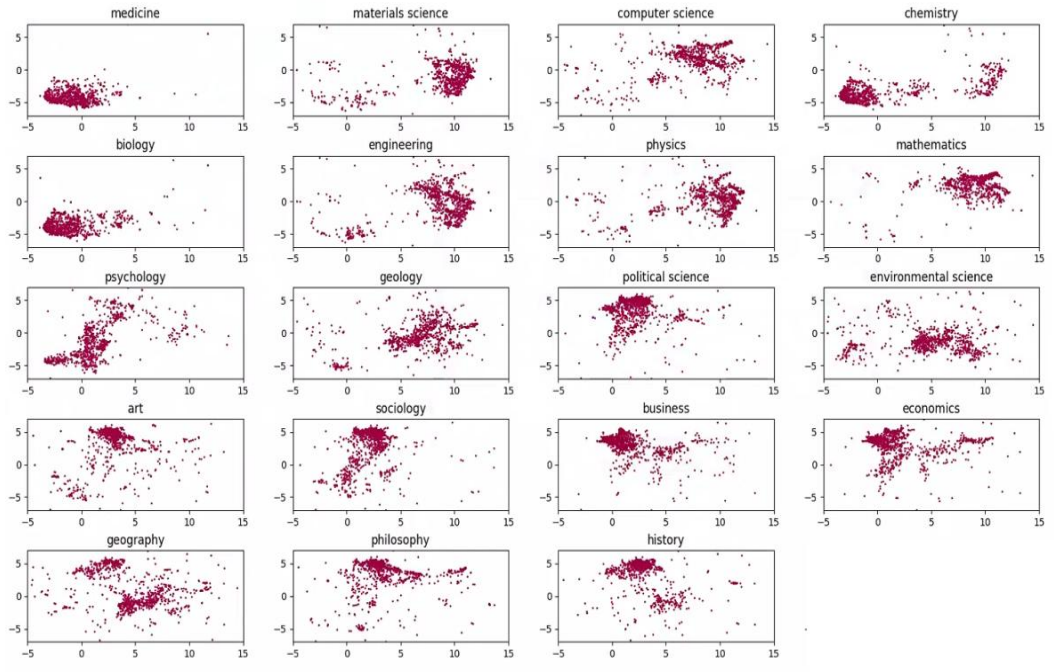


Figure 1. Semantic distribution map of various disciplines.

The vocabulary from the disciplines of medicine, chemistry, and biology exhibit a clear clustering trend in the semantic space, primarily concentrating in the lower left corner of the space. Computer science and mathematics form another concentrated area in the upper right corner. The close connection between these two disciplines may stem from their shared reliance on algorithmic thinking, logical reasoning, and theoretical modeling. Materials science, engineering, and physics are concentrated in the lower right corner of the space. This phenomenon is related to the technical and engineering methods these disciplines employ in solving practical problems.

On the other hand, the vocabulary from political science, art, sociology, business, economics, philosophy, and history is concentrated in the upper left corner of the space. These disciplines focus more on human society, culture, economy, and political phenomena, and they may have more intersections in research methods and theoretical frameworks, such as qualitative analysis, historical comparison, and critical thinking, leading to the formation of a relatively independent cluster in the semantic space.

The vocabulary from psychology, geography, environmental science, and geology is concentrated in the central region of the space. These disciplines all focus to some extent on the interaction between human activities and the natural environment. They may share common research methods and focal points in data collection, spatial analysis, and environmental monitoring, thus forming a central interdisciplinary cluster in the semantic space.

Identification of Disciplinary Diversity Types by Integrating Richness and Intersection

The results of the metrics for 19 disciplines over a 10-year period were combined and analyzed. The mean values of disciplinary richness and intersection were used as the origin, with different point shapes representing different disciplines. The horizontal axis represents disciplinary richness, and the vertical axis represents disciplinary intersection, as shown in Figure 2 below.

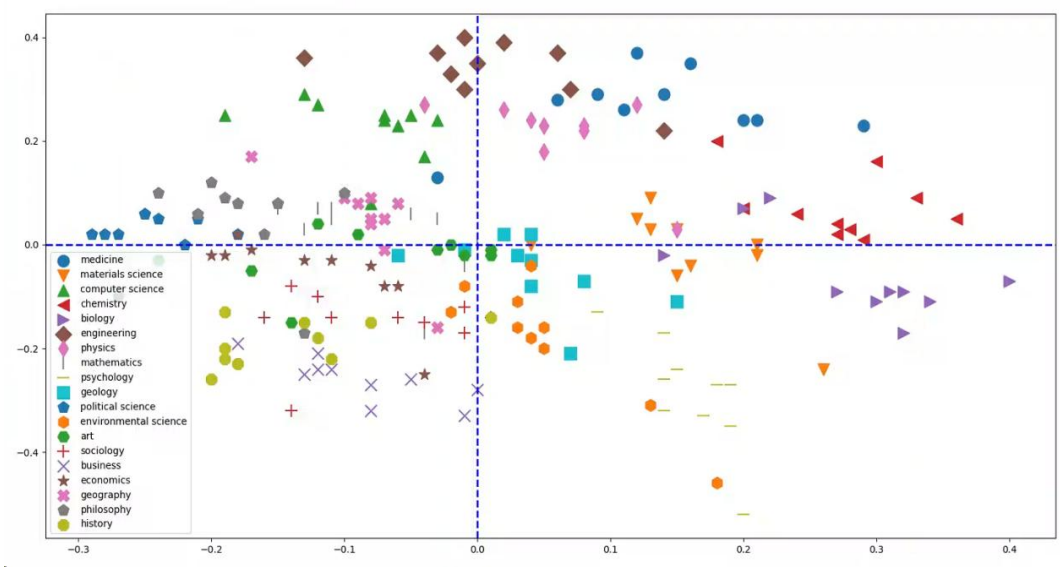


Figure 2. Classification of disciplinary diversity types.

Based on the situation in Figure 2, all points were divided into four regions according to the natural boundaries where disciplinary richness equals zero and disciplinary intersection equals zero. The division results are presented in Table 1.

Table 1. Classification results of disciplinary diversity types that integrate richness and intersectionality.

Table	*Low disciplinary richness*	*High disciplinary richness*
High interdisciplinary degree	Computer science Engineering Geography Mathematics Philosophy	Medicine Material science Chemistry Physics
Low interdisciplinary degree	History Business Political science Art Sociology Economics	Environment science Geology Psychology Biology

Analysis of Temporal Trends in Disciplinary Richness

The trends in disciplinary richness are shown in Figure 3, with high richness and high intersection in red, high richness and low intersection in green, low richness and high intersection in yellow, and low richness and low intersection in blue. Overall, the richness of most disciplines is declining, such as computer science, chemistry, and biology, while a few disciplines are experiencing an increase in richness, such as art, history, and sociology. The decline in richness for most disciplines reflects a trend towards specialization and concentration.

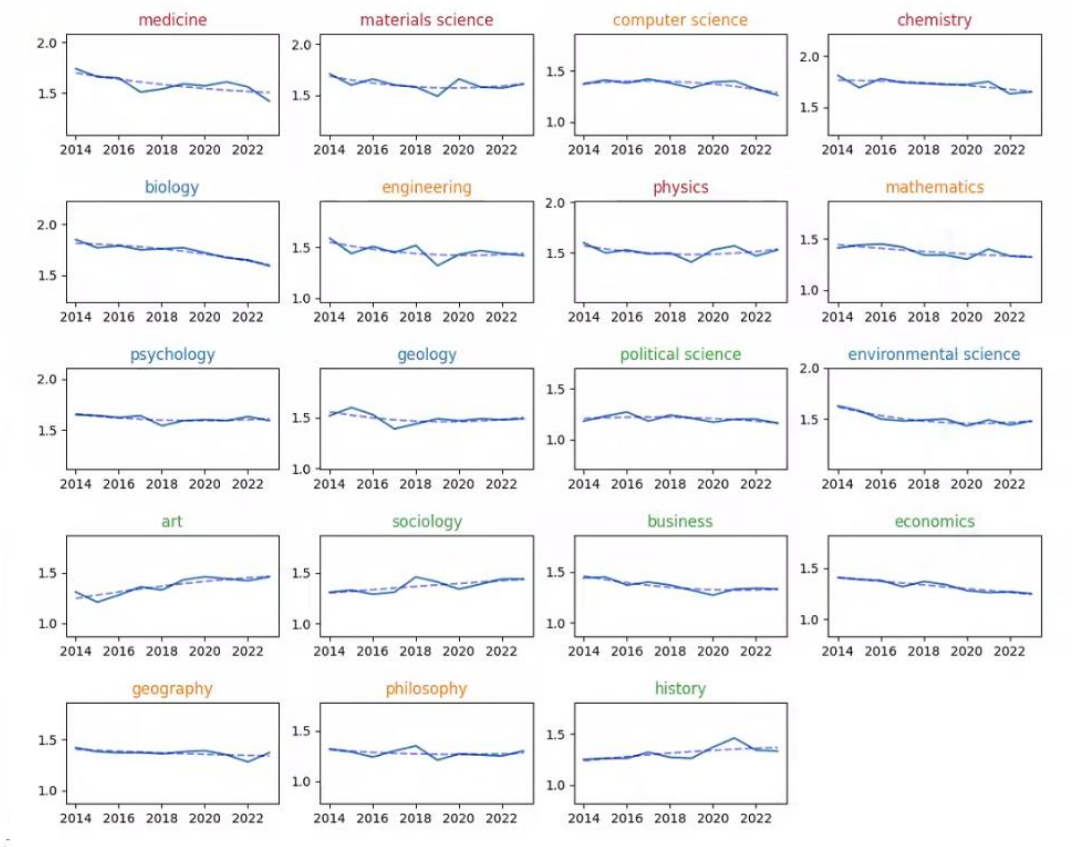


Figure 3. Time series trend chart of richness changes in various disciplines.

Analysis of Temporal Trends in Disciplinary Intersection

The trends in the intersection of various disciplines are shown in Figure 4. There is an increase in the degree of intersection for all disciplines to varying extents, reflecting a growing trend of interdisciplinary integration. As complex problems emerge, different fields begin to collaborate, sharing knowledge and technology to promote innovation and solve practical issues. This trend also reflects an increased demand for comprehensive research, leading to the gradual blurring of disciplinary boundaries and fostering the emergence of new research methods and fields.

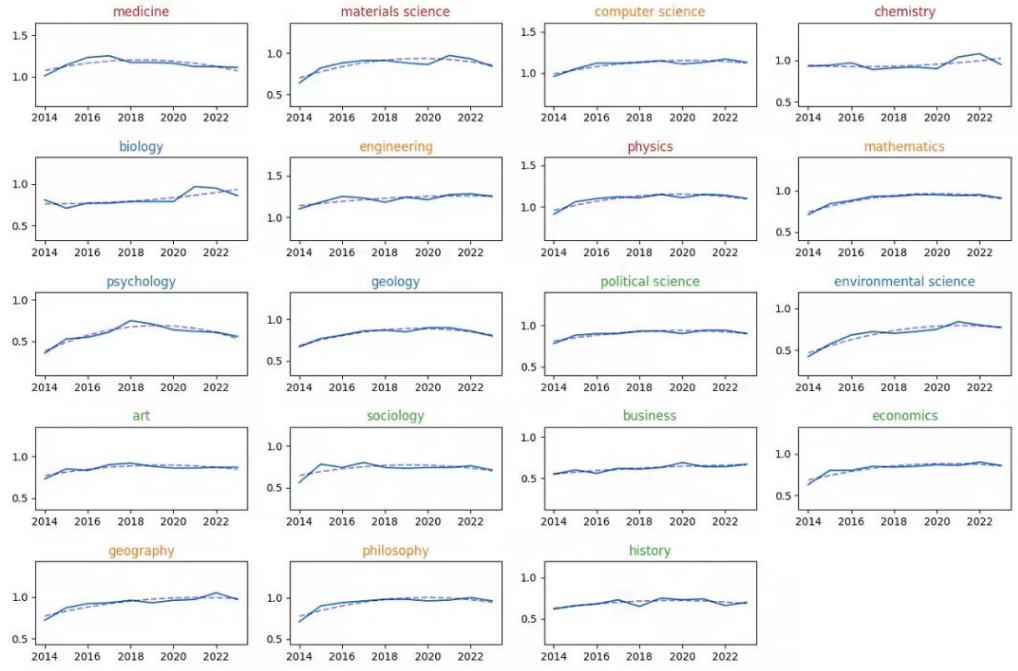


Figure 4. Trend Chart of Temporal Changes in Interdisciplinary Intersectionality among Various Disciplines.

Discussion

This paper proposes a new method for measuring disciplinary diversity based on lexical semantic analysis. Through an empirical study of articles from the OpenAlex dataset between 2014 and 2023, the effectiveness and feasibility of this method have been validated. The results indicate that this method can accurately quantify disciplinary richness and intersection from a finer-grained lexical semantic perspective, providing a new perspective for the classification and temporal change analysis of disciplinary diversity.

Despite the achievements of this study, there are some limitations. First, lexical semantic analysis relies on the quality of word embedding models and the comprehensiveness of the corpus. Imbalances in corpora across different disciplines may affect the accuracy of the measurement results. Second, this paper primarily focuses on two dimensions: disciplinary richness and intersection. Future research could consider incorporating additional dimensions, such as disciplinary balance and innovativeness, to more comprehensively reflect disciplinary diversity.

References

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*(1), 3440.

Hur, W. (2024). Entropy, heterogeneity, and their impact on technology progress. *Journal of Informetrics*, *18*(2), 101506.

Lix, K., Goldberg, A., Srivastava, S. B., & Valentine, M. A. (2022). Aligning differences: Discursive diversity and team performance. *Management Science*, *68*(11), 8430-8448.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, *118*(25), e2022340118.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the association for information science and technology*, *67*(5), 1257-1265.

https://doi.org/10.51408/issi2025_099

Single Authorship, National Co-Authorship, and International Co-Authorship in the Social Sciences and Humanities: A Multi-Dimensional Analysis of the Flemish Case

Peter Aspeslagh[1], Tim C.E. Engels[2]

[1] *peter.aspeslagh@uantwerpen.be*, [2] *tim.engels@uantwerpen.be*
Center for R&D Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract

Single-authored publications are common in the Social Sciences and Humanities, while internationally collaborative papers are less common than in other fields. Recently, we collected all author affiliations for the complete Flemish SSH research output (2012-2022). This data allow us to present a unique comprehensive analysis of national and international co-authorship in the SSH. We investigate the degree of collaboration by dividing the dataset into three distinctive subsets: single-authored publications, publications with only national collaboration and publications with international collaboration. The analysis is carried out on multiple levels: by total number of publications, by publication type, by discipline and by country of collaboration. We find that international collaboration is steadily rising, and is most common in edited volumes.

Introduction

The Humanities and Social Sciences are known to be more locally anchored than other fields of science. In the Humanities in particular, single authorship is more common than in other fields of science. In the Social Sciences, co-authorship is gradually increasing, as is co-authorship involving international co-authors (Henriksen, 2016).

In this paper, we present a unique comprehensive approach to the study of the evolution of single authorship, co-authorship at the national level, and co-authorship involving international colleagues in the Social Sciences and Humanities. For the period 2012 to 2022 we collected the affiliation data of all co-authored peer-reviewed publications included in the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (henceforth VABB, see Verleysen et al 2015). Since the VABB compiles all articles, book publications and proceedings contributions by researchers affiliated to a Flemish SSH university department, it offers comprehensive coverage of the publications in the SSH beyond that covered by international databases such as the Web of Science, Scopus, Dimensions or OpenAlex. In the case of the VABB, about half of the publications are indexed in the Web of Science (WoS), while the other half of the peer-reviewed publications (the GP publications) is included upon approval by a panel of academics appointed by the Flemish Government.

As a result of a data collection started in 2019, we have enriched all the records in the database with author affiliation data (Aspeslagh, 2024). While the affiliation data for the publications indexed in the WoS is available in WoS, author affiliations for GP publications needed to be retrieved alternatively via a multifaceted data

collection operation (matching with other sources where possible and manual look-up of all other records; registering affiliated organizations via (ROR-)identifier). This addition enables to distinguish single-authored publications from co-authored publications, which can be further subdivided into co-authored publications that involve only national collaboration (=all affiliations are Belgian) and co-authored publications that involve cross-country collaboration (=at least one of the affiliations is non-Belgian).
Based on the complete set of VABB publications for the period 2012 to 2022 we analyze:

1. The evolution of the share of single-authored publications, co-authored publications involving national collaboration, and co-authored publications involving international collaborations. We analyze this separately for each of the publication types distinguished in VABB, for the Humanities and for the Social Sciences, and for the disciplines that resort under Social Sciences and Humanities in the OECD Fields of Research and Development classification (OECD, 2025);
2. The countries that are most commonly involved in the set of internationally collaborative papers. Similarly, we analyze this separately for the Humanities and for the Social Sciences, and for the disciplines that resort under Social Sciences and Humanities in the OECD Fields of R&D classification.

The main purpose of the analysis presented in this research in progress paper is to shed light on the single authorship, national co-authorship, and international co-authorship of the Social Sciences and Humanities in Flanders, Belgium.

Method

Discipline classification

The VABB data comprehensively cover peer reviewed publications of the following types: journal articles, monographs, edited volumes, book chapters, and proceedings papers. The VABB covers the full output of all researchers affiliated to a SSH faculty at one or more of the five Flemish universities. The data are classified both according to an organizational and a cognitive classification (Guns et al 2018). In the organizational classification a discipline is assigned on the basis of the entity the author is affiliated to, independent of the content of the publication. In the *cognitive classification* the publications are assigned to OECD FoRD disciplines based on the channel in which they are published (e.g. all articles published in Scientometrics are assigned to *Computer and information sciences* and *Media and communications*).

Dataset

The initial dataset used in this analysis contains all peer-reviewed VABB publications from 2012 to 2022 (n=103,157). The dataset is available on Zenodo (https://doi.org/10.5281/zenodo.14810662). For a limited number of those publications (n=2,312, 2.2%), no affiliation data could be retrieved, resulting in a

final dataset of 100,845 publications. The number of publications per year varies between 6,975 (2012) and 10,558 (2021). In terms of affiliations and co-authorship we divide this set of publications in three:

- Single-authored publications with one or more affiliations (SAu) (n=29,830; 29.6%);
- Co-authored publications with only national (intra-Belgian) affiliations (NAf) (n=30,166; 29.9%);
- Co-authored publications with one or more international (extra-Belgian) affiliations (IAf) (n=40,849; 40.5%).

In the cognitive classification a subset of 68,558 publications are assigned to a SSH discipline. 48,376 publications are categorized as social sciences, 23,704 as humanities, with a slight overlap (multiple disciplines can be assigned). This subset, based on the cognitive classification, will be used when studying publications by discipline and comparing social sciences with humanities.

Evolution of SSH publications by collaboration type

Overall, more than 70% of the VABB publications are co-authored publications. However, the distribution over the three subsets is evolving over time. While the share of IAf is gradually increasing, the proportion of both the NAf and the SAu subsets is clearly decreasing (Figure 1). The percentage of SAu evolved from almost 33.6% in 2014 to 24.7% in 2022; NAf from 33.3% to 28.6%. Meanwhile the share a publications in international collaboration rose from one third (33.2%) of the publications in 2014, to almost half in 2022 (46.7%).

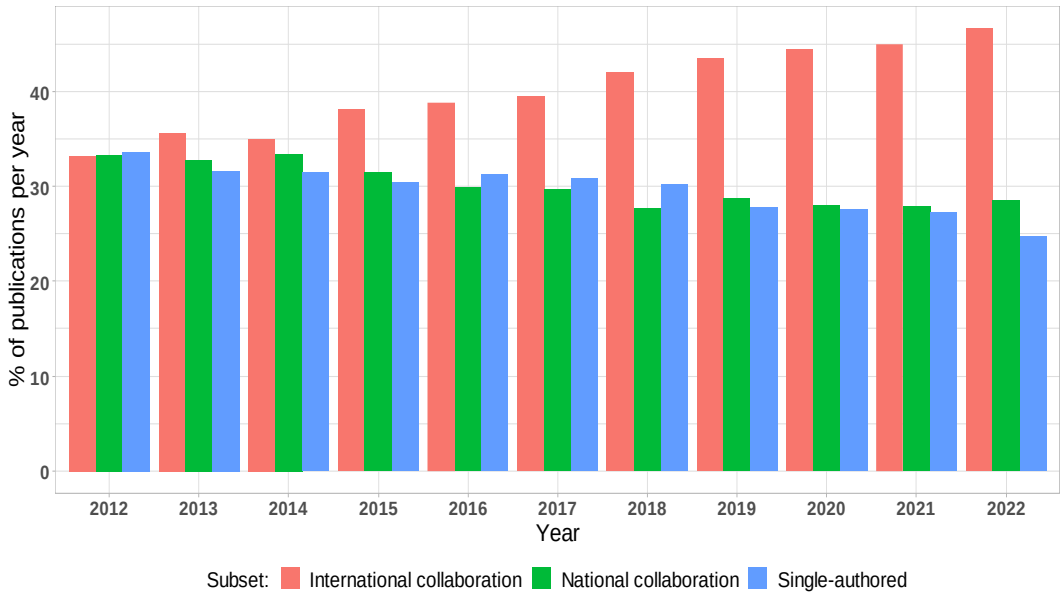


Figure 1. Collaboration in VABB per year and subset.

Publication type

VABB contains five different publication types: journal articles (74.2%), books/monographs (1.6%), edited volumes (2.2%), book chapters (17.5%) and conference proceedings (4.5%). In relative terms, IAf is the highest for edited volumes: more than half of them (58.8%) are published with a co-editor affiliated to a non-Belgian institution, a number remaining constant over time. However, the share of IAf is clearly increasing for journal articles (from 36.2% to 50.6%). Books and book chapters, in contrary, largely remain single-authored. Conference proceedings are mostly published in collaboration, evolving to a higher share of international collaboration until 2017.

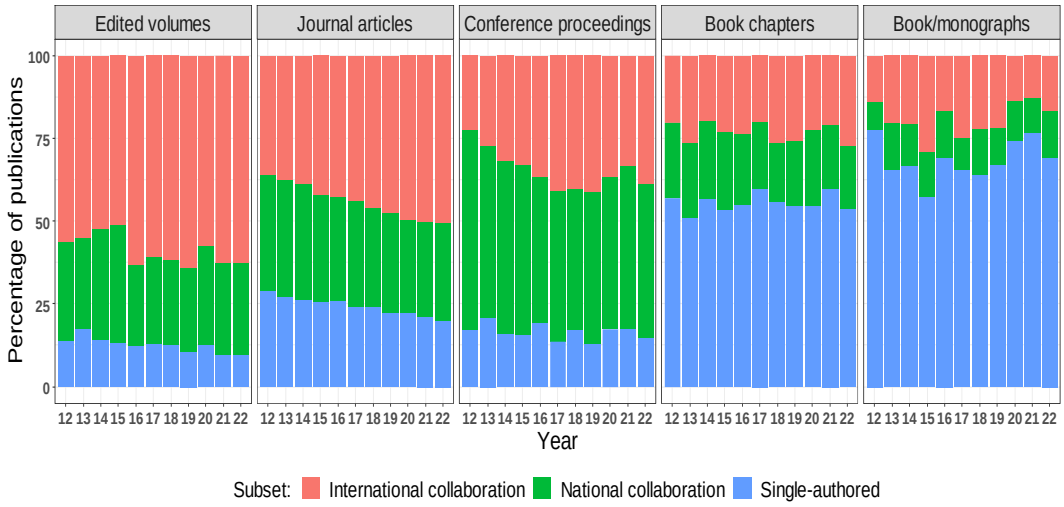


Figure 2. Collaboration by publication type.

Discipline

When reducing the full dataset to the publications that are strictly SSH according to the cognitive classification, almost two third of the humanities publications are SAu (61.0%), with both NAf and IAf near 20%. SAu, NAf and IAf are almost evenly distributed in social sciences. The absolute number of publications that are labeled as social sciences is more than double than the humanities ones (48,376 versus 23,704).

Table 1. Distribution of type of collaboration by SSH group.

	Social sciences	*%*	*Humanities*	*%*
Single-authored	15,767	32.6	14,462	61.0
National collaboration	14,851	30.7	4,668	19.8
International collaboration	17,758	36.7	4,554	19.2
Total	**48,376**	**100**	**23,704**	**100**

Zooming in on the level of individual disciplines, almost all social sciences disciplines precede their humanities counterparts concerning the degree of

international collaboration, mirroring SAu publications. Only law and, to a lesser degree, political science have a lower IAf than at least one humanities discipline.

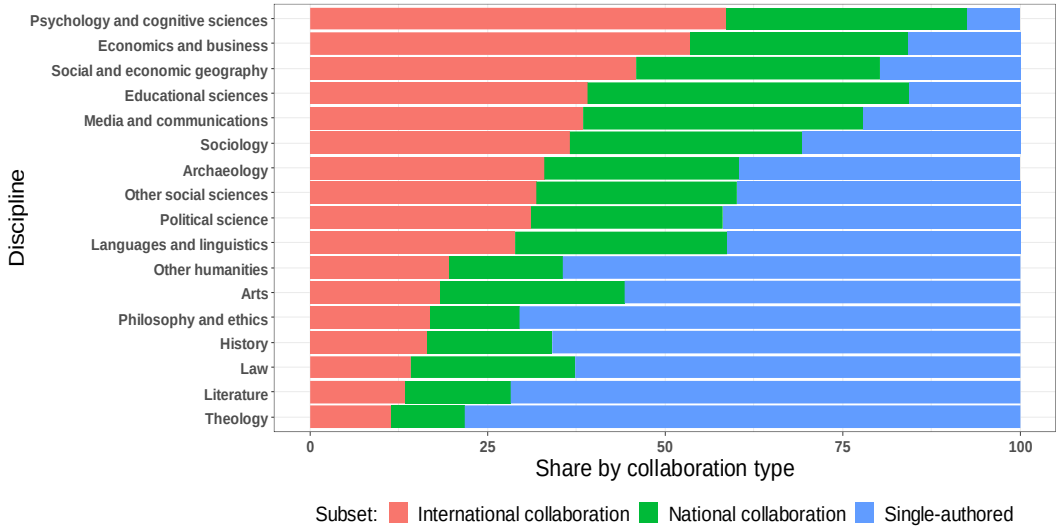


Figure 3. SSH disciplines, ordered by share of international collaboration.

Collaboration by country

During the period 2012-2022 SSH authors affiliated to a Flemish university co-authored with colleagues and scholars from 186 different countries. Collaboration most often takes place with Dutch institutions (n=12,744), followed by the United Kingdom (n=8,800), the United States (n=7,467) and Germany (n=5,913).

By disciplinary group

When switching to the strict SSH subset, the same countries are usually rounding the top 5 (NL, UK, US, DE, FR), with France sometimes being surpassed by Italy. In social sciences, the top 4 are always the Netherlands, the UK, the US and Germany, with only two years in which the internal order was different. Humanities show more variety in the order.

Except for social sciences 2021-2022, five countries always represent more than half of the publications with international collaboration. NL, GB, US and DE are always represented; a fifth country varies between FR, IT and CN.

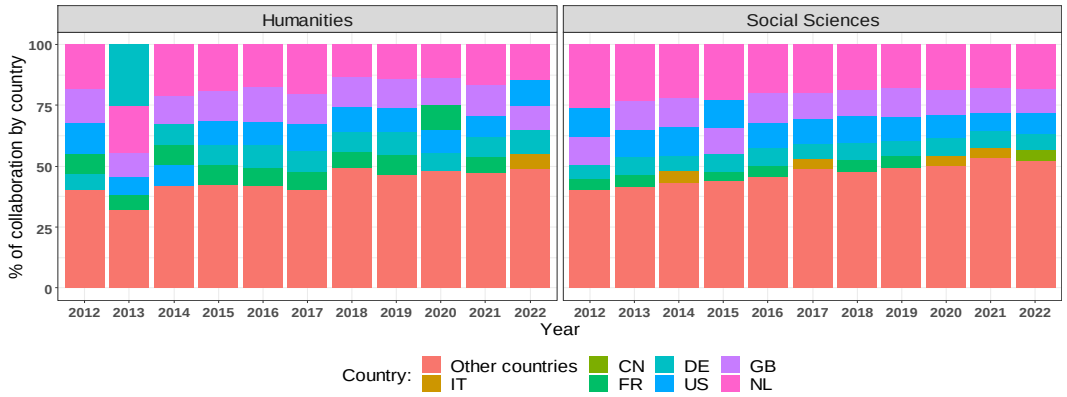


Figure 4. Percentage of affiliations per country over time (Humanities versus Social Sciences).

By discipline

The same picture appears when a distribution is made by separate discipline. In each discipline bar archaeology, the Netherlands is the country Flemish SSH authors most often collaborate with. The largest concentrations of affiliations with the top 5-countries are found in theology, history and law.

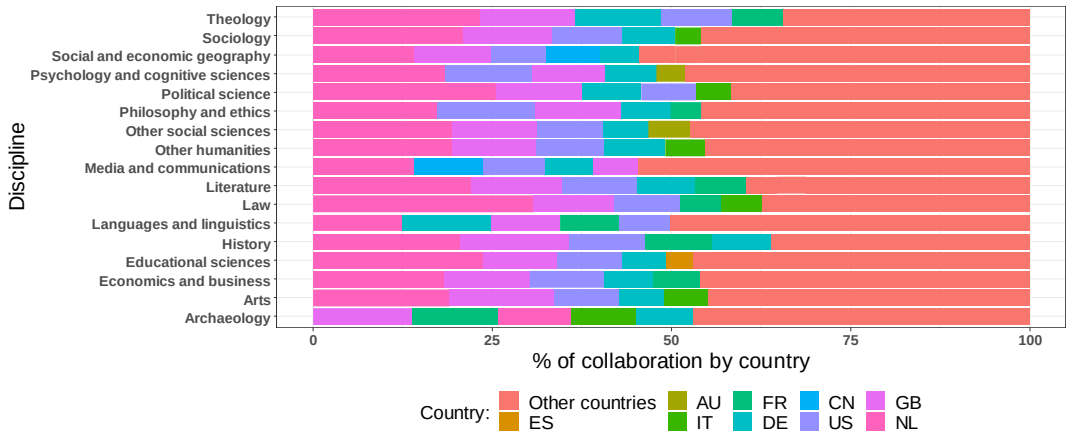


Figure 5. Percentage of affiliations per country by discipline.

Future research

During the author affiliation data collection for GP publications, organization identifiers (ROR) were assigned to each affiliation. In the near future, coding on this level will be extended to the WoS publications. This will enable us to conduct a more fine-grained analysis of organizations authors of SSH publications are affiliated with.

While registering the affiliation data, it was discovered that only two thirds of affiliated organizations were covered by ROR identifiers (Aspeslagh e.a., 2022). Therefore we are in the process of developing an extended organization database, the Flemish Organization Registry, which will include all organizations relevant for

the Flemish STIE system. It adds complementary organizations to a local ROR copy, specifically focusing on educational, research and government-related organizations. The extended organization database will be valorized in a broader context and serve other institutions and purposes.

Conclusion

The addition of author affiliation data to VABB offers a comprehensive view on collaboration in Flemish SSH research. We distinguish three subsets: single-authored publications, co-authored publications with national collaboration only and co-authored publications which include international collaboration.

In general, the share of the publications in international collaboration is rising over the time window 2012 to 2022. The share of single-authored and nationally co-authored publications is decreasing. However, when analyzing the publications by discipline, a more nuanced picture appears: in the humanities single-authored publications remain dominant, while in the social sciences single authorship, national collaboration, and international collaboration are about as common. The majority of books and book chapters remain single-authored, while journal articles, edited volumes and conference proceedings show a trend towards more collaborative publishing.

International coauthors of Flemish SSH scholars are mainly affiliated to institutions in the Netherlands, the UK, the USA, Germany and France. When studied by discipline, in most cases half of the affiliations can be assigned to a set of five countries.

References

Aspeslagh, P., Engels, T.C.E. & Guns, R. (2022). Building on ROR. Enriching and customizing multi-purpose organization databases. 26th International Conference on Science, Technology and Innovation Indicators (STI 2022), Granada. Zenodo: https://doi.org/10.5281/zenodo.6974621

Aspeslagh, P. (2024). International collaboration in SSH: looking further than the Web of Science. Research Evaluation in Social Sciences and Humanities (RESSH 2024), Galway. Zenodo: https://doi.org/10.5281/zenodo.11240671

Guns, R., Sīle, L., Eykens, J. et al. (2018). A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics* 116, 1093–1111. https://doi.org/10.1007/s11192-018-2775-x

Henriksen, D. (2016). The rise in co-authorship in the social sciences (1980-2013). *Scientometrics*, 107, 455-476. https://doi.org/10.1007/s11192-016-1849-x

OECD (2015), *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264239012-en

Verleysen, F., Ghesquière, P. & Engels, T.C.E. (2014). The objectives, design and selection process of the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW). In W. Blockmans et al. (Eds.), *Bibliometrics: use and abuse in the review of research performance* (pp. 115-125). London: Portland Press.

https://doi.org/10.51408/issi2025_114

Small Open Access Publishers: An Analysis of Visibility and Impact Patterns

Roberto Cruz Romero

cruzromero@dzhw.eu
German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)

Abstract

This research-in-progress addresses an often-overlooked dimension of scholarly publishing, namely, the transmission role of small academic publishers. Small publishers are commonly entrusted with a representation function, either serving as official outlets for specific scholarly societies or as linguistic or regional alternatives to mainstream global publications. Journals from small publishers are thus the face of these types of local organisations as they represent the material mechanisms of transmission. The transmission dimension is compounded by the visibility granted by open access, and the potential increase in impact that the latter facilitates and implies. Hence, this contribution takes visibility and impact as two fundamental characteristics of open access that small publishers commonly exploit in order to better position themselves in relation to large publishers. Using data from Scopus, Web of Science's Primary and OpenAlex, this contribution characterises the trends and patterns in publications from small publishers that primarily focus on open access as main output format. The exploratory analysis focuses on topics, linguistic and regional representation, as well as on authorship; the interplay of these dimensions sheds light on the transmission dynamics of these small outlets given that they represent focalised and/or specialised channels for authors in underrepresented regions. An underlying part of the analysis relates to the comparison of the three databases and their respective coverage of these publishers' journals.

Introduction

Research on the impact and relevance of small scholarly publishers is rather limited (Kaier & Lackner, 2019; Pinter & Magoulias, 2015; Stephen & Stahlschmidt, 2022). On one hand, there is little to observe when compared to large commercial publishers (Asai, 2020; Butler, Matthias, Simard, Mongeon, & Haustein, 2023; Larivière, Haustein, & Mongeon, 2015); i.e., the output of smaller publishers is only marginal vis-à-vis that of larger publishers. In addition, the scholarly trends regarding processing fees for open access (OA) or publication agreements are all focused on the latter group. On the other hand, some research has focused on the potential for diversification that smaller publishers can bring to mainstream fields of (meta-) study, such as bibliometrics and scientometrics (Barnes & Gatti, 2019; Giménez Toledo, Kulczycki, Pölönen, & Sivertsen, 2019). Further, other scholarly works have focused on understanding the role of smaller (usually independent publishers) regarding the dynamics of OA within the publishing landscape (Berger, 2021; Hawthorne, 2014; Ma, Buggle, & O'Neill, 2023).

Thus, this contribution seeks to enrich the limited approaches to the study of small scholarly publishing and its characteristics in relation to the latent relevance of their publications. Building on the approach by Cruz Romero et al. (2024), which focuses on the incursion and contribution of small publishers' journals into specific scientific

discourses, this research-in-progress (RP) looks at two foundational elements relating to the attributed relevance of scholarly items, namely, visibility and impact. The focus on these two dimensions seeks to problematise the interrelationships, and usually confounding proximities, of the dimensions of accessibility, costs and intellectual rights (e.g., Ball, 2016). The semantic and (very) material web of inherent issues that arise parallel to access also touch upon quality and assessment (Krüger & Hesselmann, 2020; Wiedmer, 2015). The debates surrounding these relationships tend to point toward the issues of *attention* and *prestige* (Wiedmer, 2015, pp. 150-151), directed both at the journals (as representation of the publishers) and the authors, as the bearers of intellectual weight in this regard. Thus, the interest to explore the role of the works published in these outlets regarding their impact in the scientific landscape and that of authorship in thematic (discipline) and regional distributions, particularly from the perspective of small publishers' journals.

Literature Review

As noted, the interest of this approach lies at the rather under-explored nature and characteristics of small scholarly publishers. Kaier and Lackner (2019) present some of the most notable research on this regard. Focusing explicitly on the side of small publishers, the authors exemplify the opinions and motivations of different sets of publishers, which express the diverging incentives that determine the scholarly landscape – specifically when assessing various disciplines, "possibly due to the fact that open access is already more widespread in the natural sciences than in the humanities" (p. 198). And in specific attention to the size dimension, the authors note that "smaller publishers are forced to be more 'conservative' and less innovative due to a lack of scope for investment, but this also makes them increasingly less competitive, which favours further market concentration" (p.195). This outlook marks one of (if not) the main lines of study regarding small publishers, which is their disadvantageous positions vis-à-vis larger ones.[1]

The sequential line of argument in relation to small publishers is given by the dimension of access, i.e., open access. Smaller publishers tend to be regarded as more independent and less driven by effects of market pressures and incentives (Estelle, 2021; Pinter & Magoulias, 2015). Yet, as seen, some of these elements do play a role in determining the type of publication offerings they cater. Even further, the distinction between different standards is often source of broader debates regarding rights and intellectual property – "the gratis/libre distinction, which is about rights and permissions, is not the same as the Green/Gold distinction, which is about delivery" (Ball, 2016, p. 183). Siebeck presents these debates, from the perspective of publishers, in the form of six (out of twenty-four in total) theses. The bottom line for publishers, the author argues, is the further dependency on public (or private) research funds, which will incur in indirect financing of the costs for authors. On the

[1] On the contrary, see the studies by Lariviere et al. (2015) on the so-called "oligolopoly" of academic publication, referring to the largest, most dominant scholarly publishers. Similarly, Butler et al. (2023) look at the oligopolistic dynamic of prices for OA. On the latter, Delgado-López-Cozar and Martín-Martín (2024) also contextualise and discuss the role of the business transformation of OA scholarly publishing.

authors' side, Siebeck continues, OA can re-dimension the motivations and incentives giving way to a journal submission (Siebeck, 2014, pp. 42–43).
Specifically on the latter, there is little research done on the topic, leaving great room for speculation in relation to why authors choose to publish in journals from small OA publishers. Wiedmer (2015), Siebeck (2014), Ball (2016) and others since have argued that the greatest benefit comes to readers, enjoying unrestricted access to research materials and no associated costs. Yet, the fears and strategic behaviours, some noted in Kaier and Lackner (2019, pp. 200-202), have been pointed out by Knöchelmann (2023) in that the editorial-publishing perspective carries and is often laden with many parallel debates, e.g., representation and diversity (pp. 396-397). Additionally, small publishers and, as argued by Knöchelmann, the often-confounded scholar-led publications are in constant balance between dependency and autonomous relations with the scientific community and the technical backend provided by larger publishing entities. One of the main issues at hand when dangling on between these dimensions is the visibility of published works (pp. 400-401). That is, under some circumstances, the visibility of specific symbols can be referred to as a characteristic of quality. Thus, showing off these elements becomes "relevant to encourage authors to submit, although the metrics do not allow for optimal external presentation, especially in the case of new, as yet unknown OA journals" (p. 401).
Methodologically, and at closest to the intended direction of this analysis, are the works of Stephen and Stahlschmidt (2022) and Cruz Romero et al. (2024). The first one focuses on an empirical analysis to "make evidence-based recommendations to actors in the scholarly publishing system to sustain and support the bibliodiversity offered by small publishers during the transition to OA" (Stephen & Stahlschmidt, 2022, p. 1). Stephen and Stahlschmidt's study is of particular importance since it offers the methodological backbone of this analysis, in that the same parameters and filter-matching criteria are employed here (see below). The second one more directly builds upon the analysis of small publishers, with a special focus on OA, and couples its aim with the element of *bibliodiversity* (see also Barnes & Gatti, 2019; Berger, 2021). This analysis enters into the gap that size, access and diversity leave out, thus centring on small OA publishers and the thematic patterns they seem to present in a general overview.

Data and Methods

Following the work of Stephen and Stahlschmidt (2022) and Cruz Romero et al. (2024), this contribution aims to characterise the trends and patterns of published works in journals belonging to small OA publishers. To offer precision, small OA publishers refer to a list of publishers that fulfil two criteria (one regarding size and one regarding publication licensing format). Regarding **size**, publishers are classified as small if they manage and publish ten or less journals a year *or* if, out of these journals, a total of 240 or less articles (or reviews) are published yearly (see Stephen & Stahlschmidt (2022) for methodological details on these parameters). The editorial output is coupled with the journal management aspect, given the premise that small publishers act as representatives of the academic communities they speak for and on behalf of. As the authors note in their report, this size is determined by a comparison

with the average (and median) output measured in a list of publishers identified in Crossref (pp. 6-8), given that Crossref "provides a strong, publicly available foundation for identifying small publishers" (p. 7). In this sense, the threshold harmonises different approaches and techniques.

Regarding the **access** dimension, following the same methodological cues in the abovementioned studies, there is a set threshold for the percentage of OA items published yearly in the journals managed by the publishers. Parting from the premise that the proportion of "OA documents making up typically less than 10%, or in fewer cases, more than 90% of journals' content" follows a two-sided distribution (Stephen & Stahlschmidt, 2022, p. 22). Thus, this analysis sets the threshold at 90% OA content published yearly in every journal from small publishers. This means that publishers are classified as OA if all its journals publish 90% or more of its items (per year) in some type of OA format. An important distinction here relates to the different *types of OA*; namely, if its *gold, green or hybrid*. For the purposes of this study, the distinction will not be taken into consideration given that the research interest lies at the intersection of access (in general) and broader dynamics regarding topics, linguistic and regional representation, as well as authorship.

Following these criteria, the bibliometric data comes from Scopus, Web of Science's Primary and OpenAlex databases (all queried through the infrastructure of the German Competence Network for Bibliometrics, KB – in German). Further, as mentioned, an inherent objective of the study is to compare the extent to which data is covered in all three databases and how large is the variation between each other. For that purpose, 2019 is taken as the baseline year for identifying small OA publishers,[2] which then were searched in a five-year period (2019-2023) in order to assess the trends and patterns discussed above. The focus lies not only on the mere volume of publications (which are filtered only to *articles* and *reviews* from peer-reviewed *journals*), but also on the topic classifications assigned. For this purpose both the keywords and database typologies are used, generating a differentiating parameter between the self-declared and algorithmically assigned thematic foci (e.g., see Lu et al., 2020).

Furthermore, to assess linguistic and regional representation, the metadata relating to item language and author(s) affiliation(s) are used.[3] Finally, on a meso-level perspective, the publisher location (country and city) metadata are employed to compare and draw parallels with the regional specificity or fit. This means that the analysis looks how and to what extent do the author affiliation and publisher location correspond to a match. All the steps are recoded so as to have a uniform disciplinary comparison, and this is done according to the Organisation for Economic Cooperation and Development's (OECD) Fields of Science categories (OECD,

[2] Since the following year, 2020, saw a stark downturn in the output of scholarly works largely due to the Covid-19 pandemic. This phenomenon has been broadly explored in the recent bibliometric literature.

[3] Both the topic (discipline) and the regional classifications allow for multiple coding of single items. Yet, as seen in Cruz Romero et al. (2024), the distribution does not deviate all too much from a single coding distribution.

2007).[4] Figure 1 presents a conceptual framework that links the analytical features with the variables of interest, namely, visibility and impact. From this perspective, the analysis emphasises the analytical feature as proxy elements that characterise the two dimensions of bibliodiversity discussed. These are but approximations to the empirical elements contained in the macro-level; i.e., bibliodiversity is nurtured by a diversity of topics, expressed in distinct languages, from a broad palette of regions and by a diverse group of authors. The latter elements are mediated by how visible and how impactful they are. Figure 2, in addition, summarises the methodological steps followed in the analysis.

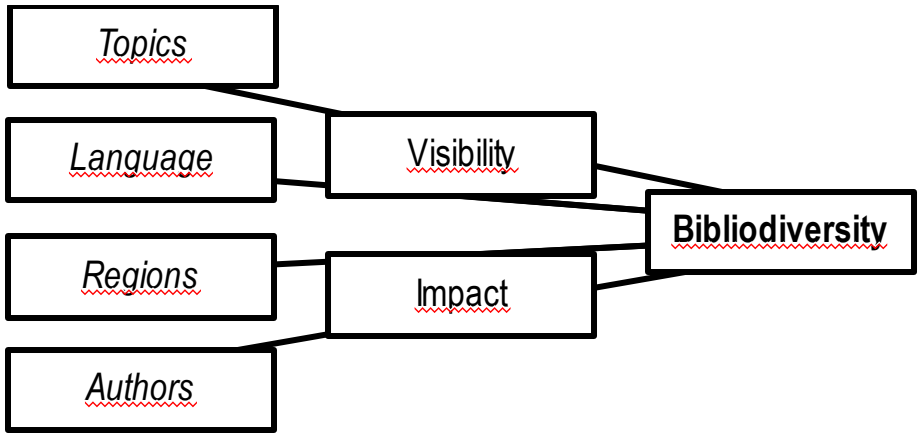


Figure 1. Conceptual Framework.

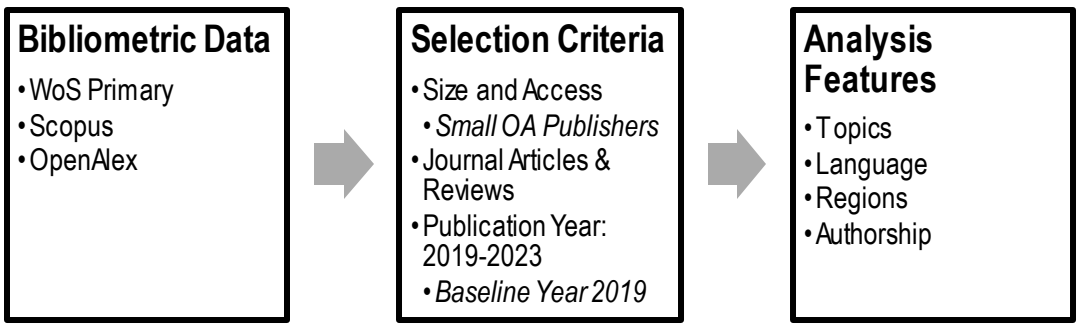


Figure 2. Methodological Flow Chart for Bibliometric Data and Analytical Framework.

Insights

An initial exploration of the data gives out a count of a total 3,145,653 *unique* items, from a total of 16,290 *unique* journals, belonging to a total of 8,691 *unique* publisher identifiers. The emphasis on the unique count is made due to the multiple counts that

[4] Social Sciences, Humanities, Agricultural Sciences, Engineering and Technology, Natural Sciences and Health and Medical Sciences.

different authoring numbers allow. That is, a single item may be single-authored; yet another item may be multi-authored. This count varies greatly within the sample, as there a mean number of authors per item of 4.1, and a median of 3 authors per item. Similarly, the collaboration of knowledge production (an intersection of physical and cognitive mobilities) is shown through the counts of institutional affiliations. In this case, there is a mean (and median) of 1.9 (1) institutional affiliations per item. In addition to these, a relevant factor for this research relates to the international affiliation differences, which can be seen in the country counts per item; again, a mean (and median) different country affiliations of 1.2 (1) per item. These figures show diverging patters in that it is inferable that a) there is a strong multi-authorship in items published in this type of outlets (journals from small OA publishers), b) the inter-institutional cooperation remains stable, yet c) scholarly cooperation seems to be nationally confined – i.e., OA journals do not seem to attract publications from international teams. This latter finding opens numerous avenues for further research. Table 1 presents a disaggregated count of these elements in the five-year period observed (2019-2023). Counterintuitively, it seems that the pandemic effect pushed the level of co-authorship (or multiple authorship) upwards, in that an evident change in the years pre- and post-pandemic occurred, driving the mean number of authors significantly (ca. 25% from the baseline 2019).[5] Furthermore, it seems that inter-institutional cooperation did increase, yet to a limited extent. It then seems as if small OA journals attract cooperation-intensive research to their pages, potentially reflecting a sector-wide phenomenon of intensive cooperative research. Yet, over 90% of items accounted for were published by teams with less than ten researchers, and the long right tail dilutes the remaining 10% 10 and over 2.000 researchers. In summary, around 20% of all items accounted are single-authored items.

Table 1. Descriptive Statistics for Bibliometric Data (Item Level).

Year	***Published Items (Unique)***	***# of Author***		***# of Institutions***		***# of Countries***	
		Mean	*Median*	*Mean*	*Median*	*Mean*	*Median*
2019	536.996	3.7	3	1.9	1	1.2	1
2020	587.664	3.9	3	1.9	1	1.2	1
2021	594.360	4.0	4	2.0	1	1.2	1
2022	579.691	4.1	4	2.0	1	1.2	1
2023	564.338	4.3	4	2.0	1	1.2	1

Moving on to the topic of linguistic and regional representation, Table 2 accounts the top ten languages and countries listed for each item. These two features differ,

[5] A small number of items account for a rather large number of authors (>2.000). This count is nonetheless correct for items authored by a multinational consortium named the ATLAS Group, with researchers from over 100 institutions worldwide. Thus, beyond one specific case in which a manual check indicated that the actual number did not match that which was registered in the database, this extreme co-authorship seems to have persisted through the pandemic years, with a series of specific publications on a narrow spectrum of the disciplinary lens – i.e., astrophysics.

nonetheless, on the unit observed, as languages refer to the item published, whilst countries are a dimension of authorship. However, put together in this sense, an intersecting dimension of authorship and mobility can be characterised. Moreover, smaller journals are expected to cater to audiences on more specific topics or lines of research that may be related with specific teams or labs. Thus, the underlying premise is that a small publisher will favour this focalised works given the disciplinary relevance that they entail for a local, national or regional expertise in the field. Moving on, expectedly, English is the outlying value in terms of publication language, yet the feature country of affiliation shows a larger variance within the sample. The underlying argument for this research-in-progress is that there are different structural and contextual incentives that determine the publication patterns in certain disciplinary groups. Moreover, the incentive structures may or may not align with international standards regarding the dominant academic discourse(s). As seen in this first stage, there appears to be a funnelling effect from different countries towards the English language – a characteristic of the contemporary academic landscape, as well as a pitfall for bibliodiversity. The premise that small OA publishers (and the journals) represent an understated research paradigm, focused more on the locally does not seem to hold. The analysis of the disciplinary distribution still follows as a next analytical step.

Table 2. Top Ten Item Languages and Author Country Affiliations.

Items Languages	***n***	***Authors Country Affiliation***	***n***
English	2.364.828	United States	352219
Portuguese	193.953	Indonesia	310860
Spanish	149.444	Brazil	289139
Indonesian	115.581	China	233026
Turkish	79.037	India	200271
Russian	67.781	Japan	188237
Japanese	34.168	Turkey	179217
French	22.268	Russia	135555
Ukrainian	17.965	United Kingdom	93880
Polish	17.063	Ukraine	87517

Limitations and Next Steps

The analysis is based on a methodological framework for identifying small publishers and their journals. In addition, the OA dimension expands (or rather further delimits) this search scope, leaving a sample of varied sources and broad disciplinary directions. The data thus only allows for *within* comparisons, i.e., comparisons between publishers and journals with the same size and access dimensions – everything else can be inferred to not belong to this category and therefore entails another set of analytical characteristics. Nonetheless, the dataset is rich and offers great insights into underexplored dimensions of academic publishing.

The next steps, not fitting in the research-in-progress format, focus on the disciplinary dissection of the features observed above (again, a sub-element seen in Stephen & Stahlschmidt, 2022; or in Severin, 2020). This approximation is relevant to correctly identify field-specific traits that may influence which type of outlet researchers tend to opt for most probably. Further, citation impact and textual analysis complement the study, providing a more comprehensive analytical framework that enters into preexisting discussions in the literature and providing fresh outlooks in these directions. Furthermore, the research will look into how smaller publishers are expected to thrive in an increasingly competitive environment of market incentives. By proxying the nature of smaller publishers (i.e., their commercial or scholarly affiliations), the discussion can be driven to the facet of sustainability, where the publish-or-perish paradigm still holds relevance. From this perspective, structural conditions (funding and affiliation) become intertwined with visibility and impact (i.e., bibliodiversity), offering further argumentative lines that this research will seek to outline.

Acknowledgments

This contribution is financed by the German Federal Ministry for Education and Research (BMBF) in its Open Access Culture funding line (project number: 16KOA014). Special thanks to Dimity Stephen for advancing the methodological framework used in this paper and Stephan Stahlschmidt for further analytical input.

References

Asai, S. (2020). Market power of publishers in setting article processing charges for open access journals. *Scientometrics*, *123*(2), 1037–1049.

Ball, D. (2016). Open Access: Effects on Publishing Behaviour of Scientists, Peer Review and Interrelations with Performance Measures. In P. Weingart & N. Taubert (Eds.), *Wissenschaftliches Publizieren*. Berlin, Boston: De Gruyter. Retrieved January 13, 2025, from https://www.degruyter.com/document/doi/10.1515/9783110448115-007/html

Barnes, L., & Gatti, R. (2019). Bibliodiversity in Practice: Developing Community-Owned, Open Infrastructures to Unleash Open Access Publishing. *ELPUB 2019 23rd edition of the International Conference on Electronic Publishing*. Presented at the nternational Con- ference on Electronic Publishing, Marseille, France: HAL. Retrieved October 9, 2023, from https://hal.science/hal-02175276

Berger, M. (2021). Bibliodiversity at the Centre: Decolonizing Open Access. *Development and Change*, *52*(2), 383–404.

Butler, L.-A., Matthias, L., Simard, M.-A., Mongeon, P., & Haustein, S. (2023). The Oligopoly's Shift to Open Access. How the Big Five Academic Publishers Profit from Article Processing Charges. *Quantitative Science Studies*, 1–33.

Cruz Romero, R., Stephen, D., & Stahlschmidt, S. (2024, September 20). *Assessing bibliodiversity through reference lists: A text analysis approach*. Zenodo. Retrieved January 13, 2025, from https://zenodo.org/records/14045592

Estelle, L. (2021). Enabling smaller independent publishers to participate in Open Access transformative arrangements. *Septentrio Conference Series*, (4). Retrieved November 7, 2022, from https://septentrio.uit.no/index.php/SCS/article/view/6220

Giménez Toledo, E., Kulczycki, E., Pölönen, J., & Sivertsen, G. (2019, December 5). Bibliodiversity – What it is and why it is essential to creating situated knowledge. *Impact of Social Sciences*. Retrieved August 15, 2023, from https://blogs.lse.ac.uk/impactofsocialsciences/2019/12/05/bibliodiversity-what-it-is-and-why-it-is-essential-to-creating-situated-knowledge/

Hawthorne, S. (2014). *Bibliodiversity: A manifesto for independent publishing* (1st publ.). North Melbourne: Spinifex Press.

Kaier, C., & Lackner, K. (2019). Open Access aus der Sicht von Verlagen: Ergebnisse einer Umfrage unter Wissenschaftsverlagen in Deutschland, Österreich und der Schweiz. *Bibliothek Forschung und Praxis*, *43*(1), 194–205. De Gruyter.

Knöchelmann, M. (2023). Herausgeberschaft und Verantwortung: Über die Un-/Abhängigkeit wissenschaftlicher Fachzeitschriften. *Bibliothek Forschung und Praxis*, *47*(2), 393–406. De Gruyter.

Krüger, A. K., & Hesselmann, F. (2020). Sichtbarkeit und Bewertung. *Zeitschrift für Soziologie*, *49*(2–3), 145–163. De Gruyter Oldenbourg.

Larivière, V., Haustein, S., & Mongeon, P. (2015). The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE*, *10*(6), e0127502.

Lu, W., Liu, Z., Huang, Y., Bu, Y., Li, X., & Cheng, Q. (2020). How do authors select keywords? A preliminary study of author keyword selection behavior. *Journal of Informetrics*, *14*(4), 101066.

Ma, L., Buggle, J., & O'Neill, M. (2023). Open access at a crossroads: Library publishing and bibliodiversity. *Insights*, *36*, 1–8.

OECD. (2007). *Revised Field of Science and Technology (FoS) Classification in the Frascati Manual* (No. DSTI/EAS/STP/NESTI(2006)19/FINAL). Working Party of National Experts on Science and Technology Indicators. Paris: Organisation for Economic Cooperation and Development. Retrieved from https://www.oecd.org/science/inno/38235147.pdf

Pinter, F., & Magoulias, M. (2015). The small academic press in the land of giants. *Insights*, *28*(3), 56–61. UKSG in association with Ubiquity Press.

Siebeck, G. (2014). 'Open Access' und offene Fragen: 24 Thesen aus verlegerischer Sicht. *Bulletin / Vereinigung der Schweizerischen Hochschuldozierenden = Association Suisse des Enseignant-e-s d'Université*, *40*(2–3), 41. ereinigung der Schweizerischen Hochschuldozierenden.

Stephen, D., & Stahlschmidt, S. (2022). *Landscape study of small journal publishers for the Knowledge Exchange Task & Finish Group for 'Small Publishers and the Transition to Open Access'*. Zenodo. Retrieved February 1, 2023, from https://zenodo.org/record/7258048

Wiedmer, H.-R. (2015). Publizieren im Zeitalter von Open Access: Die Verlagsperspektive. *Traverse: Zeitschrift für Geschichte = Revue d'histoire*, *22*(1), 147. Chronos.

https://doi.org/10.51408/issi2025_115

Structural and Institutional Determinants of Open Access Publishing: A Macro-Perspective

Roberto Cruz Romero[1], Stephan Stahlschmidt[2]

[1]*cruzromero@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)

[2]*stahlschmidt@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)
Unit of Computational Humanities and Social Sciences (U-CHASS), EC3 Research Group, University of Granada, Granada (Spain)

Abstract

The Open Access (OA) transformation is a central component of the Open Science endeavour and is frequently addressed bibliometrically due to its engagement with publication data. Whilst most studies predominantly investigate the intra-scientific effects of the OA publication model or the framework conditions for increasing the publication rate (i.e., a responsive OA environment), fewer research is focused on the economic implications (causes and consequences) of OA for the entire science system. Thus, open science and innovation systems – as material reflections or OA consolidation – represent contexts co-determined by precise political frameworks, where policies and mandates can have direct impacts on individual challenges with clear societal implications, such as free access to scientific literature. In line with this pivotal role for knowledge transfer, the political level is a fundamental dimension with which to assess the OA patterns and the subsequent adoption of broader open science policies. Hence, both academic and regulatory dimensions of scientific production find themselves at the same crossroads, highlighting the institutional and systemic roles of that both funders and researchers play. Methodologically, we approach the relationship between policy frameworks and open science from the access dimension. We use data from three major bibliometric databases: WoS, Scopus, and OpenAlex. We are interested in observing the regional distinctions for the OA trend and try to identify geographically bound tendencies in the OA publication landscape. For that reason, we further match country-level administrative data with more specific "academic space" indicators, thus trying to uncover structural conditions that hinder or promote OA adoption. In line with recent explorations, we find that OA shows a stagnating pattern, whilst "closed" research has seen an uptake. OA costs appear flexible for richer countries than for lower income countries, which depend on a larger extent on fee-waver programs for access to read and publish in OA journals. We extend the analysis to an inferential approach through a nested logit regression, a type of multinomial logistic model to observe the probability of choosing to publish in open access compared to closed access. We discuss policy implications for the publishing landscape, as well as for the innovation-oriented scientific system.

Introduction

The Open Access (OA) transformation is a central component of the Open Science endeavour and is frequently addressed bibliometrically due to its engagement with publication data. A host of studies predominantly investigate the intra-scientific effects of the OA publication model or the framework conditions for increasing the publication rate; more broadly, studies focus on a responsive OA environment within the scientific system. Economic implications of the OA model are predominantly discussed concerning the market power of the five largest academic publishers

(Larivière, Haustein, & Mongeon, 2015), while financially weaker actors outside the scientific system, such as small and medium-sized enterprises with research interests, are initially excluded from the reuse of scientific content due to high subscription costs (Bryan & Ozcan, 2021). The latter dynamic creates a centrifugal tendency that pushes academic outputs into the realm of commodification through price-setting and access gatekeeping.

Thus, open science and innovation systems represent contexts co-determined by precise political frameworks, where policies and mandates can have direct impacts on individual challenges with clear societal implications, such as free access to scientific literature. Accordingly, the observation of the various political measures to promote the OA transformation appears practice relevant. From a political perspective, openness correspondingly refers to the conditions necessary for creating innovation incentives. In line with this pivotal role for knowledge transfer, the political level is a fundamental dimension with which to assess the OA patterns and the subsequent adoption of broader open science policies.

Nevertheless, the open knowledge generation and diffusion has to be supported by clear policy goals that complement and expand the scholarly and economic systems. As stated by Bai (2014), and reiterated by Sá and Grieco (2016), OA to research outputs needs an institutional backing so as to effectively link diverse productive actors and generate virtuous systems of research and development, as well as innovation.

Hence, OA – as a part of the open science and open innovation systems – can effectively have direct impacts on problems with great societal ramifications whilst, that is, being driven by precise policy frameworks. For example, Sá and Grieco (2016) present a still reverberating discussion about the role of open data as a result, but also as a driving force for policymaking, promoting transparency and accountability of the research output. In this case both outputs and primary data constitute the base upon which policy debates are conducted, contrasting academic, administrative, and economic perspectives.

The Institutional Perspective on Open Access

Open access has become the beacon of hope for many of its advocates. The *unpaywalled* access[1] to research should benefit scholars in disadvantageous economic circumstances (Knöchelmann, 2021), as costs for reading are eliminated through the offsetting of these by, mostly, author-sided costs. The system of OA publishing then has become dependent on large-scale agreements between publishers and research institutions (or the funding bodies supporting these). The largest European nations, for example, committed to adopting open science and OA publishing practices since the Budapest and Berlin declarations.[2]

[1] Here unpaywalled refers only to the cost-free (subscription or pay-to-read) access to scholarly research, and not to the platform with the same name.

[2] Further initiatives have been developed regarding access to publicly-funded research, such as the Helsinki Initiative for Multilingualism (Federation Of Finnish Learned Societies, Information, Publishing, Universities Norway, & European Network For Research Evaluation In The Social Sciences And The Humanities, 2019), the Vienna Principles for Scholarly Communication (Kraker

This evolving dynamic has led to the development of consequent policy frameworks, such as the European Commission's official endorsement of the San Francisco Declaration on Research Assessment (DORA), or the German Research Foundation's positioning regarding academic publishing, and as of late, the push to consolidate the Coalition for Advancing Research Assessment (CoARA). Alongside the institutional consolidation of the OA narrative, the publishing landscape has grown at a much faster pace, leaving funding bodies with many challenges to the guidelines and positions they have vis-a-vis research output. For instance, the dispersion of the OA modalities is the most telling sign of a rapidly changing environment. Beall characterises the OA publishing movement as something "concerned more with the destruction of existing institutions than with the construction of new and better ones" (2015).

This destruction is conceived within the scope of the licenses used to characterise OA (which are commonly colour-coded) and differentiate between direct (e.g., the *gold* format) and indirect routes. The latter, i.e., the *green* model, allows authors to make their works available before, during, or after the journal publication, mainly through personal or institutional repositories.[3] The *bronze* route is every sense like the green, but with a key difference regarding rights and permissions, where the bronze option can be imprecise. Transformative agreements (such as DEAL) are part of a wider scope of publishing mechanisms in which authors can choose to pay so-called article processing charges (APCs) in order to "open" their research in, predominantly, non-OA journals.

Data and Methods

To approach the relationship between policy frameworks and open science, we look firstly at the access dimension. For that, we focus on the open access information available in the bibliometric data infrastructure of the German Bibliometrics Competence Network (KB - in German). As seen in the previous plot, to frame our approach of OA growth and stagnation, we downloaded data from three major bibliographic databases: WoS, Scopus, and OpenAlex. We compare various snapshots and highlight the need for a systematised and complete dataset (see Figure 1). However, we base our exploration on the August 2024 snapshot of the OpenAlex database. We must note that bibliographic data suffers from a time-sensitive correction (as seen in Figure 1). This dynamic introduces some level of imprecision in the data exploration as for the overall counts, which we want to make noted.

Our focus is only on *articles* published between 2014 and 2023, for a time series subset of ten years. Moreover, we have two main characterisations regarding the OA information, a) as open vs. closed, and b) sub-divided into the colour categorisation

et al., 2016), the Jussieu Call for Open Science and Bibliodiversity ('Jussieu Call for Open Science and Bibliodiversity', 2017) and, more recently, the Barcelona Declaration on Open Research Infrastructure (Barcelona Declaration on Open Research Information, Kramer, Neylon, & Waltman, 2024).

[3] Specific right allocations and permissions are dependent on the editorial rights used by each publisher, and are commonly associated to a Creative Commons license (CC. See, e.g.: https://creativecommons.org/licenses/by-sa/4.0/deed.en.

described above. We delve deeper into the publisher dimension of the OA landscape and characterise two distinct dimensions of publishers: size and OA distribution. The former approach is determined based on the methodology proposed by Stephen and Stahlschmidt (2022), in which publishers are categorised according to their yearly outputs, whether in terms of articles or journals published. Seen in Figure 1, the OA momentum has been systematically driven by these larger publishers, so it becomes relevant to identify their overall weight.

Then, we are interested in observing the regional distinctions for the OA drive and the geographically bound trends in the publication landscape. So, we use of the World Bank's open data repository, and its classification in both regions and income groups. The former is based strictly on geographic bases, and hence the groupings observed. As hinted, we also look at the income level differences provided by the same dataset, dividing countries between high, upper-middle, lower-middle, and low income.

A further level of analysis is focused on the academic spaces of countries in both income and regional classifications. For this approach, we make use of the Varieties of Democracy's (V-Dem) *Academic Space* indicators, conceptualised as proxies for academic freedom (Coppedge et al., 2024). We look here into four specific dimensions of this set of measurements: freedom to research and to teach, to exchange and disseminate, to act as critics, and the institutional autonomy (in relation to the latter, the extent to which institutional autonomy is granted or, on the contrary, hindered). To fully characterise the academic space, also from an economic perspective that relates to the income and regional classifications, we complement with data on the share of public spending directed at research and development activities. Finally, in the context of academic space, we look at the Research4Life initiative, and its classification of countries that are eligible for special funding for access to specific publisher portals. This classification is highly correlated with both a geographic and income level typology, so we explore their relation further. We run logistic regressions on the publication trends in OA for an unbalanced panel of countries in the 2014-2023 timeframe using the `mlogit` package (Croissant, 2020) in the R framework.

The following results are meant as guiding insights for a deeper discussion on the role of OA publishing regarding open science and open innovations. We present preliminary data explorations that allow us to build a comprehensive perspective on, firstly, the apparent stagnation of OA and, secondly, the structural determinants that characterise it. Unless stated otherwise, all data is presented in the time frame mentioned and processed from the OpenAlex snapshot, always focusing on the access dimension.

Results

The preliminary results of the data analysis shown in the subsequent sections are an exploration of the distinct features that we seek to emphasise. Thus, the insights drawn from these approximations serve as an entrance point to a larger analytical framing and should serve as preview for the subsequent inferential analysis.

Open Access Development

Figure 1 displays the relative counts of OA publications in each of the databases considered. As seen, the trends in all three follow a most similar direction, i.e., a constant growth of OA over time (with steeper increments in OpenAlex), yet a clear slowdown (even stagnation) towards the last two periods (2022 and 2023). This behaviour illustrates our interest in the structural and institutional determinants of OA publishing, given that they constitute an underlying condition for individual researchers to opt for this publication path – that is, in addition or despite institutional policies towards closed access publications.

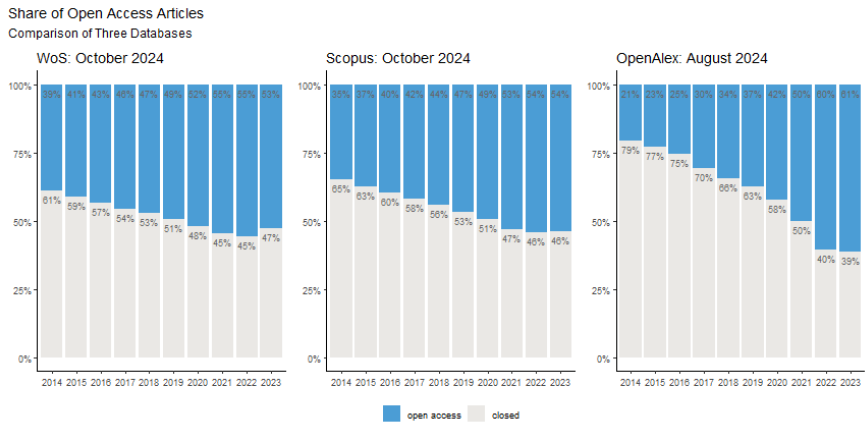


Figure 1. Open Access Trend Comparison in Web of Science, Scopus and OpenAlex (2014-2023).

Figure 2 presents a more detailed outlook of the OA trends in relation to the colour coding commonly used to indicate the corresponding licensing arrangements (see Beall, 2015).The first striking detail is the 5-p.p. change in the diamond OA articles published during the period of analysis, which reflects a limited uptake of this form of publishing standard. For clarification, *diamond* refers to those articles which licensing not only involves a creative commons copyright (CC), but crucially eliminates any payment from either side (authors and readers). This can be understood as the truly open standard. All other colour (licensing) schemes involve some other form of payment and/or limitation on the free availability of manuscripts (or data).

Looking to complementarily classify more structural dynamics that shape the allocation of resources for research, we use data from the WB that allow us to match countries to their respective income level (Figure 3). This typology is based on general thresholds of gross domestic product (GDP) that distinguish four groups: 1) High income (HI), 2) upper middle income (UMI), 3) lower middle income (LMI), and 4) lower income (LI) countries. Methodologically, we note that not all countries listed on the WB data are present in the bibliographic data from OpenAlex. Presumably, not every country that is listed had an academic affiliation which produced an article included in the database (a double contingency that limits the

scope of this study). We complement this perspective with academic space indicators.

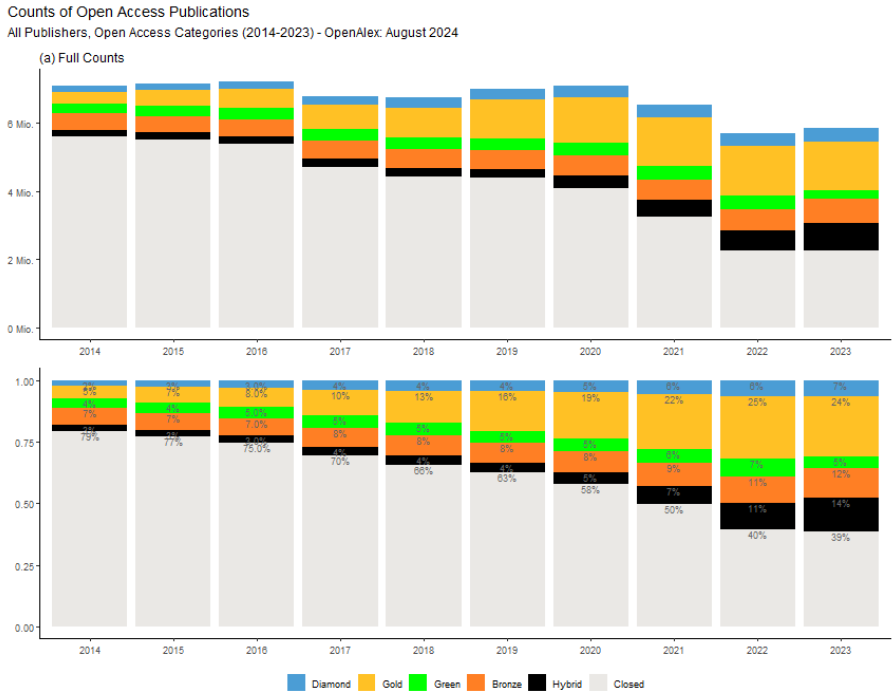


Figure 2. Open Access Trend Breakdown in OpenAlex (2014-2023).

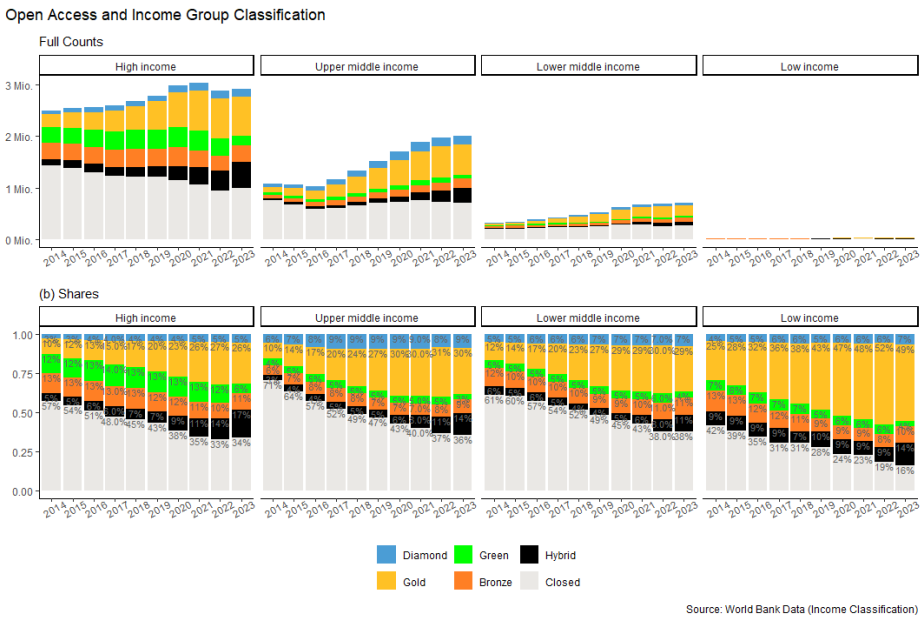


Figure 3. Open Access Breakdown by Income Classification (2014-2023).

To that end, we use the Varieties of Democracy (V-Dem) project's ***Academic Freedom Index***, or AFi. This index is composed of a series of disaggregated indicators that make up the "academic space", which characterises more thoroughly the *de jure* and *de facto* gaps in expert based and collected factual data (see Spannagel, Kinzelbach, & Saliba, 2020). We note that this index is not flawless and recognise the methodological shortcomings this type of construct may imply. We do not make use of the AFi as such, but rather we look at four precise indicators that make up the academic space characterisation included in the V-Dem dataset.[4] The four dimensions we consider are: 1) Research and teaching, 2) academic exchange and dissemination, 3) academics as critics, and 4) institutional autonomy.

Finally, we match the data with the R4L initiative dataset. The organisation funds and subsidises access to scholarly content by cataloguing countries according to their scores in different indices that relate to human development. Countries are then placed in eligibility groups to provide "institutions in low-and middle-income countries with online access to academic and professional peer-reviewed content" (see https://www.research4life.org/about/). In this sense, the lists of the R4L initiative are a proxy of reduced access to research outputs and, therefore, of analytical relevance to match with the WB and OpenAlex data.

Logistic Regression

Looking to tie together the analytical elements hitherto discussed, we now turn to an inferential analysis based on a multinomial logistic regression (MLR). We approach the analysis from this perspective given that the response variable we are interested in has the characteristic of being nominal (i.e., neither ordinal nor numeric), and we are focused on the probabilistic changes from one category to the other. In this sense, we take the bibliometric data grouped at the country level (based on author affiliation),[5] and count the total number of publications, as well as the distribution of these in OA categories – our base category, however, is closed access. To include all necessary features of interest, we match the data with the previous dataset already presented and discussed, i.e., with economic and administrative data from the WB (GDP and income groups), from V-Dem (AFi indicators), and from R4L initiative's group classification.

The data has an unbalanced panel structure, since we have the timeseries of ten years (2014-2023) with distinct number of observations in each of the categories of OA, our response variable. We group by country and year so as to aggregate the data to a macro-level of analysis.

The crux of the modelling structure, however, lies at the nature of the conditions for estimating a multinomial regression. Since we are dealing with alternatives of similar conditions, i.e., with most of the options that could be grouped under the larger nest of OA, we recur to the implementation of a nested logit model. To that end, we recode the data to, firstly, differentiate between the nest categories, i.e., open v.

[4] We use version 14 of the country/year dataset, in which the academic space indicators are found in section 3.15.4 of the Codebook.

[5] We do not estimate fractional counting when assigning country affiliation distributions. For matters of exploration and direct interpretation, we proceed with full counting of authors.

closed access. Then, we estimate the change between nested sub-categories (diamond, gold, green, bronze, and hybrid), since we part from the characteristic that "some alternatives may be joined in several groups" (Croissant, 2020, p. 21). Following this rationale, the nested logit model relaxes a key assumption regarding the cross-elasticity of alternatives (this is, that "the introduction of any new mode or the improvement of any existing mode will affect all other modes proportionally" (Forinash & Koppelman, 1993, p. 98) – with modes referring here to the alternatives). In this sense, the condition of independence of irrelevant alternatives (IIA) is relaxed and allows for correlated error terms of the groups of alternatives; put otherwise, this estimation technique parts from the premise that not only case-specific characteristics determine the choice, but also alternative-specific characteristics have a probabilistic weight for maximising the utility in the decision-making. Figure 4 (an adaptation from Forinash & Koppelman, 1993) details the nested estimation approach. As seen in the figure, we differentiate between these characteristics, viz. case (individual or, in this scenario, country-level) characteristics and alternative characteristics. The former are the same across alternatives, whilst the latter vary across choices.

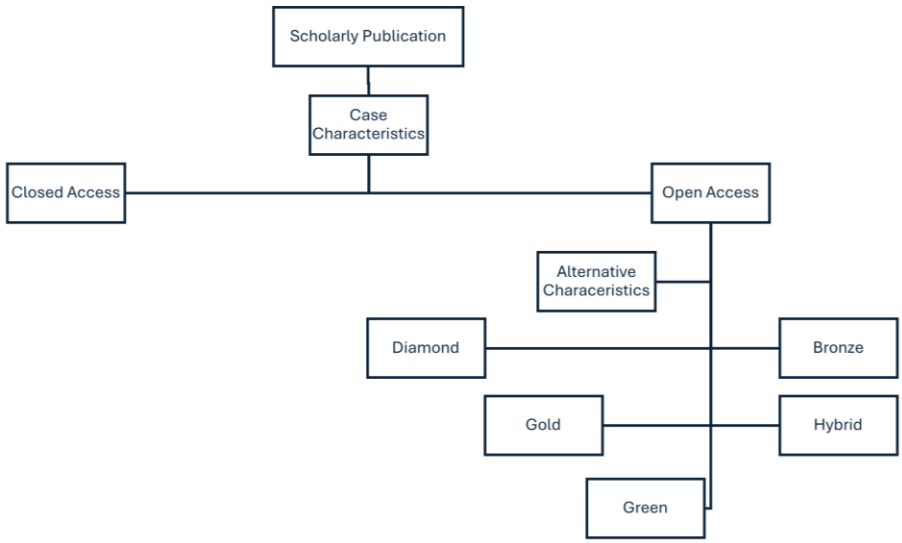


Figure 4. Example of Nested Estimation Approach.

Limitations and Next Steps

The pending analysis of the regression estimates follows a data intensive computation with the `mlogit` package in R (Croissant, 2020). The structure of the data (long format, i.e., short *T* and large *N*) creates computational demands that make it necessary to recur to a special server. The latter's capacity varies according to overall shared usage, which caused unforeseen delays in the process of preparing this submission. Hence, the inferential analysis will be further developed and presented at the conference in full scope.

Acknowledgments

This contribution is financed by the German Federal Ministry for Education and Research (BMBF) in its Open Access Culture funding line (project number: 16KOA014).

References

Alatas, S. F. (2003). Academic Dependency and the Global Division of Labour in the Social Sciences. *Current Sociology*, *51*(6), 599–613. SAGE Publications Ltd.

Bai, C. (2014). Promoting research and innovation with open access. *National Science Review*, *1*(3), 323.

Barcelona Declaration on Open Research Information, Kramer, B., Neylon, C., & Waltman, L. (2024, April 16). Barcelona Declaration on Open Research Information. Zenodo. Retrieved June 18, 2024, from https://zenodo.org/doi/10.5281/zenodo.10958522

Beall, J. (2015). What the Open-Access Movement Doesn't Want You to Know. *Academe (May-June 2015: 'I'll Tell It and Think It and Speak It and Breathe It')*, *101*(3). Retrieved June 17, 2024, from https://www.aaup.org/issue/may-june-2015-ill-tell-it-and-think-it-and-speak-it-and-breathe-it

Berger, M. (2021). Bibliodiversity at the Centre: Decolonizing Open Access. *Development and Change*, *52*(2), 383–404.

Bryan, K. A., & Ozcan, Y. (2021). The Impact of Open Access Mandates on Invention. *The Review of Economics and Statistics*, *103*(5), 954–967.

Cope, B., & Kalantzis, M. (2014). Changing knowledge ecologies and the transformation of the scholarly journal. In B. Cope & A. Phillips (Eds.), *The Future of the Academic Journal (Second Edition)* (pp. 9–83). Oxford: Chandos Publishing. Retrieved November 7, 2022, from https://www.sciencedirect.com/science/article/pii/B9781843347835500021

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Bernhard, Mi., et al. (2024). V-Dem [Country–Year/Country–Date] Dataset v14. Varieties of Democracy (V-Dem) Project. Retrieved from https://www.v-dem.net/data/the-v-dem-dataset/country-date-v-dem-v14/

Croissant, Y. (2020). Estimation of Random Utility Models in *R*: The **mlogit** Package. *Journal of Statistical Software*, *95*(11). Retrieved February 4, 2025, from http://www.jstatsoft.org/v95/i11/

Delgado-López-Cozar, E., & Martín-Martín, A. (2024). La ruta de oro de la publicación científica: Del negocio de las revistas a las revistas negocio: La fuente del negocio editorial: el negocio bibliométrico de la evaluación científica. *Revista Mediterránea de Comunicación*. Retrieved February 22, 2024, from https://www.mediterranea-comunicacion.org/article/view/26763

Federation Of Finnish Learned Societies, Information, T. C. F. P., Publishing, T. F. A. F. S., Universities Norway, & European Network For Research Evaluation In The Social Sciences And The Humanities. (2019). Helsinki Initiative on Multilingualism in Scholarly Communication, 621757 Bytes. figshare.

Forinash, C. V., & Koppelman, F. S. (1993). Application and Interpretation of Nested Logit Models of Intercity Mode Choice. *Transportation Research Record*, (1413). Retrieved February 4, 2025, from https://trid.trb.org/View/385097

Johansson, M. A., Reich, N. G., Meyers, L. A., & Lipsitch, M. (2018). Preprints: An underutilized mechanism to accelerate outbreak science. *PLOS Medicine*, *15*(4), e1002549.

Jussieu Call for Open Science and Bibliodiversity. (2017). . Retrieved March 18, 2024, from https://jussieucall.org/jussieu-call/

Khanna, S., Ball, J., Alperin, J. P., & Willinsky, J. (2022). Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies*, *3*(4), 912–930.

Khoo, S. Y.-S. (2019). Article Processing Charge Hyperinflation and Price Insensitivity: An Open Access Sequel to the Serials Crisis. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *29*(1), 1–18.

Knöchelmann, M. (2021). The Democratisation Myth: Open Access and the Solidification of Epistemic Injustices. *Science & Technology Studies*, *34*(2), 65–89.

Knöchelmann, M. (2023). Herausgeberschaft und Verantwortung: Über die Un-/Abhängigkeit wissenschaftlicher Fachzeitschriften. *Bibliothek Forschung und Praxis*, *47*(2), 393–406. De Gruyter.

Kraker, P., Dörler, D., Ferus, A., Gutounig, R., Heigl, F., Kaier, C., Rieck, K., et al. (2016). The Vienna Principles: A Vision for Scholarly Communication in the 21st Century. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, *69*(3–4), 436–446.

Larivière, V., Haustein, S., & Mongeon, P. (2015). The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE*, *10*(6), e0127502.

Leonelli, S. (2022). Open Science and Epistemic Diversity: Friends or Foes? *Philosophy of Science*, *89*(5), 991–1001.

Picarra, M. (2015). *Open Access To Scientific Information: Facilitating Knowledge Transfer And Technological Innovation From The Academic To The Private Sector* (Briefing Paper No. 611742). PASTEUR4OA (p. 9). Brussels: European Commision. Retrieved from Zenodo.

Sá, C., & Grieco, J. (2016, September). Open data for science, policy, and the public good. *Review of Policy Research*. WILEY.

Spannagel, J., Kinzelbach, K., & Saliba, I. (2020). *The Academic Freedom Index and Other New Indicators Relating to Academic Space: An Introduction* (No. 26). Users Working Paper (p. 28). Gothenburg: V-Dem Institute, University of Gothenburg. Retrieved from https://www.v-dem.net/media/publications/users_working_paper_26.pdf

Spiekermann, K. (2020). Epistemic network injustice. *Politics, Philosophy & Economics*, *19*(1), 83–101. SAGE Publications.

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature*, *495*(7442), 426–429. Nature Publishing Group.

https://doi.org/10.51408/issi2025_136

Unraveling the Driving Factors of Team Performance: The Impact of Team Composition and Collaboration Relationships on Project Teams

Ruinan Li[1], Tingcan Ma[2], Beibei Sun[3], Yuzhuo Wang[4]

[1]*liruinan@ahu.edu.cn*
School of Management, Anhui University, Hefei 230601 (China)
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

[2]*matc@mail.whlib.ac.cn*
National Science Library (Wuhan), Chinese Academy of Sciences, Wuhan 430071 (China)

[3]*sunbeibei@ncwu.edu.cn*
Department of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046 (China)

[4]*wangyuzhuo@ahu.edu.cn*
School of Management, Anhui University, Hefei 230601 (China)

Abstract

With research project teams increasingly serve as engines of scientific breakthroughs, understanding the factors driving their performance is essential and urgent. This study examines the effects of team composition (team size, gender diversity), internal collaboration (network density), and external collaboration (domestic, international, and industry partnerships) on team productivity and team impact. Using a sample of 206 research projects funded by the National Natural Science Foundation of China (NSFC), we employ Ordinary Least Squares (OLS) regression and Lindeman-Merenda-Gold (LMG) analysis to identify the most influential factors of team performance. Our results indicate that domestic and international collaborations significantly drive team productivity, while international collaboration also plays a key role in enhancing team impact. An internally dense network negatively affects team productivity but positively contributes to team impact, underscoring the nuanced nature of collaborative dynamics. In contrast, team size and gender diversity are not significant drivers for either outcome. Overall, these findings enrich a multidimensional understanding of the complex relationships between team characteristics and project performance, and offer actionable insights for managers, policymakers, and funders seeking to optimize team performance.

Introduction

In an era of rapidly evolving scientific and technological advancements, the complexity of research problems often exceeds the capacity of any single individual or discipline. Consequently, collaborative project teams have emerged as core vehicles for driving innovation (Liu, Wang, & Yang, 2025). Such teams integrate diverse expertise and resources, enabling them to address multifaceted challenges more effectively than individual researchers. Funded research projects, in particular, have been linked to a greater number of publications and high-impact outputs (Langfeldt, Bloch, & Sivertsen, 2015). As the impact and innovation performance of these project teams garner increasing attention, how the compositional features of

teams affect their performance has become a central concern for researchers. For example, recent studies underscore the importance of team size and gender diversity as critical factors influencing scientific team performance (Tang, Shi, Wu, & Li, 2023; Zhang et al., 2024; Zhao et al., 2024). Another portion of research has focused on collaboration relationships among the research and innovative activity (Whittington, 2018), and found that both internal collaboration networks and external partnerships (domestic, international, or cross-sector) significantly shape research outcomes.

Despite extensive research on team-based innovation, relatively few studies focus on research project teams (Liu et al., 2025), and many of those investigate only one dimension, such as team composition or a single facet of collaboration. Consequently, it remains unclear how multiple team characteristics collectively affect team outcomes and, crucially, the relative impact each dimension exerts. This gap is particularly relevant for research project teams, where insights into the degree of influence from team composition and collaboration relationships have important implications for improving productivity and generating high-impact publications. Motivated by this gap, this study addresses two key questions: (1) Do team composition, internal collaboration, and external collaboration significantly affect performance in research project teams? and (2) To what extent do these factors influence team performance, and which factor has the most significant impact? By addressing these questions, the study aims to offer an evidence-based perspective on how a multidimensional view of team characteristics can help optimize research outcomes. The findings will offer valuable insights for project managers and policymakers, especially in the context of research-based projects funded by institutions such as the NSFC.

Research Hypotheses

Drawing from the perspectives of team composition and collaboration relationships, we identify several key factors influencing team performance. These include team size, gender diversity, collaboration network density, as well as domestic, international and industry collaboration.

Team composition and team performance. Regarding team size, multiple studies reveal a curvilinear or inverted U-shaped pattern linking team expansion to productivity and impact. For instance, Zhao et al. (2024) found that although increasing the number of "thought leaders" can enhance team performance initially, excessive expansion reduces disruptive potential. Similarly, Tang et al. (2023) detected that while co-authorship generally elevates citation impact, indiscriminate growth of teams may not be prudent, echoing Zhu et al.'s evidence of diminishing returns beyond an optimal threshold (Zhu, Liu, & Yang, 2021). Moreover, Perović, Radovanović, Sikimić, and Berber (2016) found that smaller research teams often prove more productive. Turning to gender diversity, gender diversity in scientific teams can lead to better outcomes. Teams with gender heterogeneity can produce higher-quality publications. Zhang et al. (2024) demonstrated that moderate inter-gender collaboration promotes greater disruptive knowledge relative to single-gender teams. However, some research findings do not always support the

conclusion that gender diversity can definitely improve team performance. Wang, Wu, and Li (2024) detected an inverse U-shaped link between the proportion of women scientists and citation impact. Additionally, Sandström and Van Den Besselaar (2019) found no performance penalty for gender diversity. Therefore, the study proposes the following hypotheses:

H1a: Team size positively influences team productivity.

H1b: Team size positively influences team impact.

H2a: Gender diversity positively influences team productivity.

H2b: Gender diversity positively influences team impact.

Internal collaboration and team performance. Internal collaboration networks significantly shape research outcomes. For instance, Shalley and Perry-Smith (2008) utilized network analysis to discover that teams characterized by strong relational ties, as well as those with weaker ties, exhibited the highest levels of creativity. Singh, Tan, and Mookerjee (2011) distinguished between internal and external cohesion within team networks, and noted a positive correlation between internal cohesion and team productivity. Meanwhile, Ma, Ba, Zhao, and Sun (2023) highlighted that balanced "social capital" within and beyond the team, and collaboration features that combine internal cohesion with external linkages support high-quality scientific breakthroughs, suggesting the value of flexible and well-configured network relationships. Therefore, the study proposes the following hypotheses:

H3a: Internal network density positively influences team productivity.

H3b: Internal collaboration network density positively influences team impact.

External collaboration and team performance. In terms of external collaboration, studies consistently show that international collaborations often yield higher citation counts. Specifically, papers resulting from international or multinational partnerships generally receive more citations compared to those involving only domestic collaborations (Persson, 2010). Likewise, Abramo, D'Angelo, and Costa (2019) observed that research teams with higher levels of internationalization enjoy increased citation probabilities. Additionally, it has been proven that research teams engaging in industry-university-research collaboration are highly effective in promoting innovative research (Gray & Sundstrom, 2010; Skute, Zalewska-Kurek, Hatak, & de Weerd-Nederhof, 2019). Some research points out that this kind of collaboration has a significant positive impact on the research productivity of university research teams (Chen & Wang, 2021). Therefore, the study proposes the following hypotheses:

H4a: Domestic collaboration positively influences project team productivity.

H4b: Domestic collaboration positively influences project team impact.

H5a: International collaboration positively influences project team productivity.

H5b: International collaboration positively influences project team impact.

H6a: Industry collaboration positively influences project team productivity.

H6b: Industry collaboration positively influences project team impact.

Data and methods

This study focuses on the Innovative Research Groups funded by the NSFC. The sample consists of 206 research teams working on projects funded by NSFC from 2007 to 2024 (completed projects). These teams are involved in eight scientific fields.

We initiated the data collection process by extracting the number of members and their names from the official project information provided by the NSFC. Subsequently, we conducted an in-depth online investigation. By exploring academic profiles on institutional websites, professional networking platforms, and other reliable online sources, we were able to determine the gender of each team member. This information was then used to calculate the team size (the total number of members) and team gender diversity (using the Blau index).

This study utilized the Web of Science database as the data source for retrieving the publications of research teams. Based on the unique project grant numbers, we searched the Web of Science database. Subsequently, the retrieved publication data was imported into the ItgInsight software (http://itginsight.com/) for author cleaning. After the cleaning process, the co-authorship matrix of team members was exported. Based on the co-authorship matrix, the R programming language was utilized to construct the co-authorship network of team members. Through this network, internal collaboration network density is calculated. Finally, the search results from the Web of Science database were linked to the Incites database. In the Incites database, data on the indicators of Web of Science Documents and Category Normalized Citation Impact (CNCI), and indicators related to industry collaborations, domestic collaborations, and international collaborations were obtained.

Based on the collected data, we utilized the descriptive statistics, correlation analysis, regression analysis to analyze the impact of various factors on team performance. Furthermore, the Lindeman-Merenda-Gold (LMG) method was used to assess the relative importance of each independent variable in explaining the variance of the dependent variable.

Results

Descriptive statistics. After acquiring the data, the variables were measured, and thus the descriptive statistical analysis was performed. The results are shown in Table 1.

Correlation Analysis. The Pearson correlation analysis was conducted on variables including Project_Duration (PD), Team_Size (TS), Gender_Diversity (GD), Network_Density (ND), Domestic_Collaborations (DC), International_Collaborations (ITC), Industry_Collaborations (IDC), Web_of_Science_Documents (WOSD) and Category_Normalized_Citation_Impact (CNCI). The results are presented in Figure 1.

Table 1. Descriptive statistical analysis of variables.

	Sample size	Min.	Max.	Mean	Std_dev	Median
Project_Duration	206	3	9	6.379	1.401	6
Team_Size	206	4	18	9.204	1.601	10
Gender_diversity	206	0	0.5	0.197	0.171	0.18
Network_Density	206	0.244	1.667	0.6	0.281	0.533
Domestic_Collaborations	206	2	939	101.99	94.99	81.5
International_Collaborations	206	0	356	65.267	54.005	51.5
Industry_Collaborations	206	0	81	4.903	10.586	1
Web_of_Science_Documents	206	6	1231	267.544	202.923	217
Category_Normalized_Citation_Impact	206	0.319	5.842	1.504	0.769	1.349

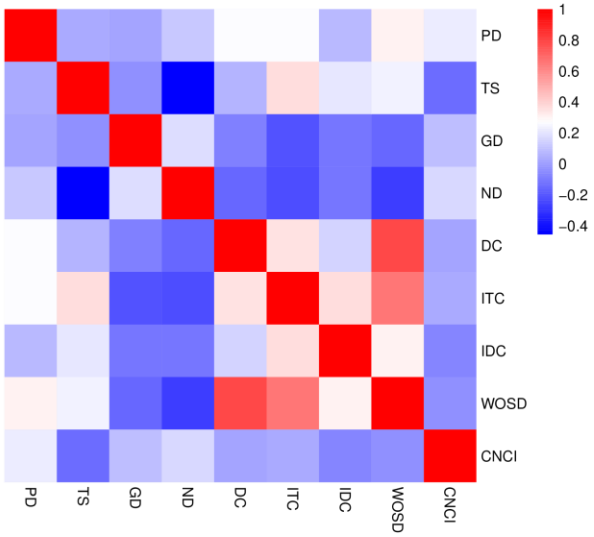


Figure 1. Correlation analysis of variables.

Regression analysis. This study employed the regression analysis to examine the relationships between a set of independent variables and two dependent variables: the Web of Science Documents and Category Normalized Citation Impact. Additionally, two control variables were included: Project_Duration (PD) and Research_Field (RF). Table 2 shows the regression results. As shown in Table 2, domestic collaborations and international collaborations stand out as significant drivers of team productivity, while international collaborations have notable effects on team impact. Notably, a denser internal network negatively influences team productivity but positively associates with team impact. By contrast, team size and gender diversity do not demonstrate significant effects on either outcome. Therefore, hypotheses of H3b, H4a, H5a and H5b are supported. And H1a, H1b, H2a, H2b, H3a, H4b, H6a and H6b are not supported.

Lindeman-Merenda-Gold (LMG) analysis. Table 3 presents the value of relative importance metrics in the LMG analysis for both models, where the Domestic_Collaborations shows the highest explanatory power, indicating that domestic collaborations contribute the most to the variation in Web of Science documents. The International_Collaborations also ranks high. For team impact, the control variables of Research_Field and Project_Duration have the highest explanatory contribution. Other variables show relatively low values. Additionally, the proportion of variance explained by the two models is 84.5% and 20.9%, respectively.

Table 2. Brief regression analysis results.

Terms	*Web_of_Science_Documents*				*Category_Normalized_Citation_Impact*			
	Coef	*Std. Err*	*t*	*p*	*Coef*	*Std. Err*	*t*	*p*
Team_Size	-2.96	4.155	-0.712	0.476	-0.045	0.051	-0.873	0.383
Gender_Diversity	23.665	35.925	0.659	0.51	0.412	0.402	1.024	0.306
Network_Density	**-52.394**	**19.189**	**-2.73**	**0.006****	**0.418**	**0.21**	**1.991**	**0.046***
Domestic_Collaborations	**1.094**	**0.127**	**8.589**	**0.000*****	-0.001	0.001	-1.01	0.312
International_Collaborations	**1.437**	**0.151**	**9.517**	**0.000*****	**0.002**	**0.001**	**2.092**	**0.036***
Industry_Collaborations	-0.184	0.435	-0.424	0.672	0	0.004	0.055	0.956
R^2	0.845				0.209			
R^2(within)	0.833				0.151			

Note: * p<0.05, ** p<0.01, *** p<0.001

Table 3. Relative importance in the LMG analysis for both models.

	Web_of_Science_Documents	*Category_Normalized_Citation_Impact*
Research_Field	0.174	0.116
Project_Duration	0.040	0.040
Team_Size	0.014	0.012
Gender_diversity	0.006	0.006
Network_Density	0.025	0.022
Domestic_Collaborations	0.348	0.001
International_Collaborations	0.213	0.009
Industry_Collaborations	0.025	0.002

Conclusion and discussion

This study aims at understanding how team composition, internal collaboration and external collaboration affect the performance of research project teams. Regarding team productivity, domestic and international collaborations have been identified as significant positive drivers. For team impact, international collaborations have a notable positive effect. Notably, the internal network density shows a negative impact on team productivity but a positive association with team impact. However, team size and gender diversity do not show statistically significant effects on either team productivity or impact. Additionally, the LMG analysis reveals that domestic collaborations have the highest explanatory power for team productivity, followed by international collaborations. For team impact, the control variables of RF and PD

have the highest explanatory contributions. Most of these results are understandable. Both internal and external collaboration relationships have a significant impact on team performance. However, team size and gender diversity are not statistically significant, perhaps due to the interplay of other contextual variables like research discipline and project duration.

Our findings contribute to the existing literature on team-based innovation, especially in the context of research project teams. Previous studies often focused on single-dimension investigations. Our multi-dimensional analysis shows that different aspects of team characteristics have distinct effects on team performance. These findings can provide actionable insights for project managers and policymakers, especially in the Innovative Research Groups funded by the NSFC.

One major limitation of this study is that the model for team impact only explains 20.9% of the variance, indicating that there are many unaccounted-for factors. This suggests that future research should explore additional variables that may influence team impact. Moreover, the role of moderating and mediating variables in the relationships between team composition, collaboration relationships and team performance should be further explored. Furthermore, to gain a more profound understanding of how to achieve the success of project teams, future research should conduct causal inference analysis, such as the application of propensity score matching (PSM).

Acknowledgments

This work was supported by the Hubei Key Laboratory of Big Data in Science and Technology (Wuhan Library of Chinese Academy of Science) (Grant Nos. E4KF011001) and the National Social Science Fund of China (No.24CTQ027).

References

Abramo, G., D'Angelo, C., & Costa, F. (2019). The correlation between the level of internationalization of a country's scientific production and that of relevant citing publications. Paper presented at the 17th International Conference on Scientometrics and Informetrics, ISSI 2019-Proceedings.

Chen, A., & Wang, X. (2021). The effect of facilitating interdisciplinary cooperation on the research productivity of university research teams: The moderating role of government assistance. Research evaluation, 30(1), 13-25.

Gray, D. O., & Sundstrom, E. (2010). Multi-Level Evaluation of Cooperative Research Centers: Bridging between the Triple Helix and the Science of Team Science. Industry Higher Education, 24(3), 211-217.

Langfeldt, L., Bloch, C. W., & Sivertsen, G. (2015). Options and limitations in measuring the impact of research grants—evidence from Denmark and Norway. Research evaluation, 24(3), 256-270.

Liu, Z., Wang, C., & Yang, J. (2025). The effects of scientific collaboration network structures on impact and innovation: A perspective from project teams. Journal of Informetrics, 19(1), 101611.

Ma, Y., Ba, Z., Zhao, H., & Sun, J. (2023). How to configure intellectual capital of research teams for triggering scientific breakthroughs: Exploratory study in the field of gene editing. Journal of Informetrics, 17(4), 101459.

Perović, S., Radovanović, S., Sikimić, V., & Berber, A. (2016). Optimal research team composition: data envelopment analysis of Fermilab experiments. Scientometrics, 108, 83-111.

Persson, O. (2010). Are highly cited papers more international? Scientometrics, 83(2), 397-401.

Sandström, U., & Van Den Besselaar, P. (2019). Performance of Research Teams: results from 107 European groups. Paper presented at the ISSI.

Shalley, C. E., & Perry‐Smith, J. E. (2008). The emergence of team creative cognition: the role of diverse outside ties, sociocognitive network centrality, and team evolution. Strategic Entrepreneurship Journal, 2(1), 23-41.

Singh, P. V., Tan, Y., & Mookerjee, V. (2011). Network Effects: The Influence of Structural Social Capital on Open Source Software Projects. Management Information Systems Quarterly, 35(4), 813-829.

Skute, I., Zalewska-Kurek, K., Hatak, I., & de Weerd-Nederhof, P. (2019). Mapping the field: a bibliometric analysis of the literature on university–industry collaborations. The Journal of Technology Transfer, 44(3), 916-947.

Tang, X., Shi, W., Wu, R., & Li, S. (2023). The expansion of team size in library and information science (LIS): Is bigger always better? Journal of Information Science, 01655515231204800.

Wang, Y., Wu, Q., & Li, L. (2024). Examining the influence of women scientists on scientific impact and novelty: insights from top business journals. Scientometrics, 1-26.

Whittington, K. B. (2018). A tie is a tie? Gender and network positioning in life science inventor collaboration. Research Policy, 47(2), 511-526.

Zhang, M.-Z., Wang, T.-R., Lyu, P.-H., Chen, Q.-M., Li, Z.-X., & Ngai, E. W. (2024). Impact of gender composition of academic teams on disruptive output. Journal of Informetrics, 18(2), 101520.

Zhao, Y., Wang, Y., Zhang, H., Kim, D., Lu, C., Zhu, Y., & Zhang, C. (2024). Do more heads imply better performance? An empirical study of team thought leaders' impact on scientific team performance. Information processing & management, 61(4), 103757.

Zhu, N., Liu, C., & Yang, Z. (2021). Team size, research variety, and research performance: do coauthors' coauthors matter? Journal of Informetrics, 15(4), 101205.

https://doi.org/10.51408/issi2025_105

Unveiling Tortured Phrases in Humanities and Social Sciences

Alexandre Clausse[1], Fidan Badalova[2], Guillaume Cabanac[3], Philipp Mayr[4]

[1]*alexandre.clause@univ-tlse3.fr*
Université de Toulouse, IRIT UMR 5505 CNRS, 118 route de Narbonne, 31400 Toulouse (France)

[2] *fidan.badalova@gesis.org*, [4]*philipp.mayr@gesis.org*
GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Köln (Germany)

[3]*guillaume.cabanac@univ-tlse3.fr*
Université de Toulouse, IRIT UMR 5505 CNRS, 118 route de Narbonne, 31400 Toulouse (France)
Institut Universitaire de France (IUF), Paris (France)

Abstract

A small amount of unscrupulous people, concerned by their career prospects, resort to paper mill services to publish articles in renowned journals and conference proceedings. These include patchworks of synonymized contents using paraphrasing tools, featuring tortured phrases, increasingly polluting the scientific literature. The Problematic Paper Screener (PPS) has been developed to allow articles (re)assessment on PubPeer. Since most of the known tortured phrases are found in publications in science, technology, engineering, and mathematics (STEM), we extend this work by exploring their presence in the humanities and social sciences (HSS). To do so, we used the PPS to look for tortured abbreviations, generated from the two social science thesauri ELSST and THESOZ. We also used two case studies to find new tortured abbreviations, by screening the Hindawi EDRI journal and the GESIS SSOAR repository. We found a total of 32 multidisciplinary problematic documents, related to Education, Psychology, and Economics. We also generated 121 new fingerprints to be added to the PPS. These articles and future screening have to be investigated by social scientists, as most of it is currently done by STEM domain experts.

Introduction

Scientific research is a cumulative process involving rigorous reporting of the experiments carried out and the observed results, in a textual or visual form, typically presented as scientific articles. These findings are then submitted to an editorial board or group of peers in order to be published, after a peer review process. The so-called 'publish or perish' paradigm implies to publish as many articles as possible in reputable journals to have the most impactful research in a given scientific community (Biagioli & Lippman, 2020). The publication pressure on individual researchers leads a small number of unscrupulous people, concerned about their career prospects, to resort to falsification, fabrication, and plagiarism.

Plagiarism can be disguised using paraphrasing tools such as spinners (e.g., SpinBot) to rephrase textual contents, including established scientific concepts. When a scientific concept is paraphrased, at least one of its terms gets replaced by a synonym, making it nonsensical regarding the associated discipline. For example, a 'convolutional brain organization' is a spun version of the 'convolutional neural network', which is known as a tortured phrase (Cabanac, Labbé & Magazinov, 2021). They are assumed to be evidence of paper mill products, a company that sells fake scientific articles, ensuring that they will be published in known journals, by manipulating editorial and publishing processes (Abalkina *et al.*, 2025; Nazarovets,

2024). The consequence of such behavior is twofold: (1) some (sensitive) research relies on these articles, making them unreliable, and (2) this leads to a major trust issue in science.

To address this, the Problematic Paper Screener (PPS) was launched in 2021; it builds upon the Dimensions bibliometric database full-text search, and allows (re)assessing questionable articles on the PubPeer platform (Cabanac, Labbé & Magazinov, 2021; Barbour & Stell, 2020). As of today, the 'tortured' detector flagged more than 18k scientific articles containing at least 5 different tortured phrases, only 2.9k of which have been retracted. However, this is tedious work, requiring several domain experts to read each article to update the PPS fingerprints list (i.e., known tortured phrases). Moreover, not all the disciplines have yet been considered, as they are mostly related to science, technology, engineering, and mathematics (STEM) studies.

We propose to extend this work and search for the presence of tortured phrases in humanities and social sciences (HSS) articles and developing guidelines to raise the awareness of the scientific community regarding such fraudulent content.

Motivation

Tortured phrases are widely polluting the scientific literature (Van Noorden, 2023). Some of them are claiming false information (Texeira da Silva, 2021), even in sensitive research related to COVID-19 (Texeira da Silva, 2023). These publications are unreliable and causing major trust issues in science. Moreover, some of these articles are used as the foundations for other studies, and false information and errors spreads. These articles featuring unreliable references are known as 'feet of clay' publications (Cabanac, 2024).

In 2022, a post publication peer-review (PPPR) approach using the Problematic Paper Screener (PPS) and PubPeer has been proposed as part of an initiative to decontaminate the scientific literature (Cabanac, 2022). It focuses on two main tasks: (1) investigate the suspect paper and (A, B) extract all the problematic content to check if (C) it has been commented on PubPeer, and (2) (re)assess it using (A) the PPS and (B) PubPeer in order to discuss its content, as depicted in Figure 1.

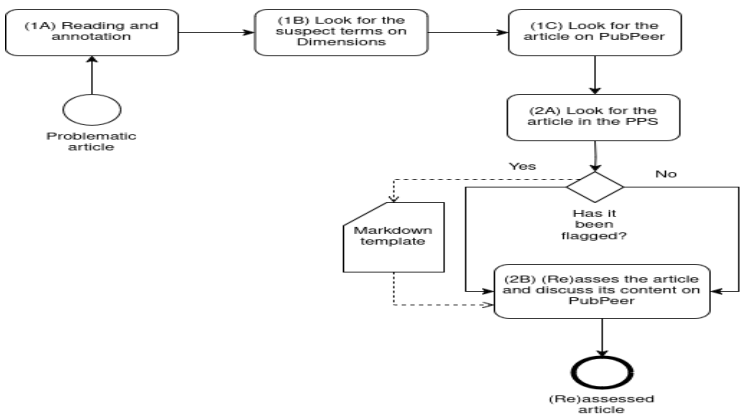


Figure 1. The research quality insurance workflow, describing how an analyst looks for tortured phrases in an article flagged by the PPS and (re)assesses it on PubPeer.

As of today, the PPS features 6,545 fingerprints (i.e., known tortured phrases) which are used to screen the scientific literature by querying the Dimensions database, as their documents textual contents are indexed in. Since the 'tortured' detector already found tortured phrases in 18,582 document, most of them are related to STEM studies, as depicted in Figure 2.

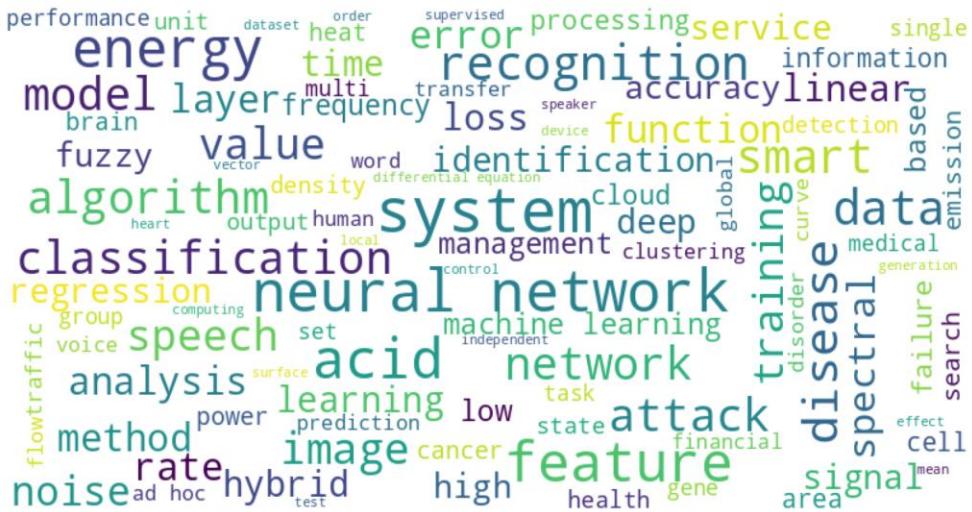


Figure 2. Word cloud of terms and phrases from the fingerprints list.

However, exploring further this list of fingerprints, it appears that it also contains 42 tortured phrases related to HSS (e.g., 'electronic democratic framework', the tortured phrase for 'electronic voting machine'). Moreover, we found irregularities while looking at the term 'computer-assisted telephone interviewing (CATI)', where the word 'telephone' has been replaced by 'phone' in several scientific articles, making the abbreviations mismatching its developed form.

We studied the European Language Social Science Thesaurus (ELSST) and the GESIS Thesaurus for the Social Sciences (THESOZ) thesauri that showed subtleties to be taken into account. As an example, 'civil war' and 'internal war' are genuine concepts, although we first thought one was the tortured version of the other. This suggests that the HSS vocabulary is less standardized as the STEM one.

Materials and method

We adopted a systematic approach to identify bogus text by focusing on tortured abbreviations (Clausse, 2023; O'Grady, 2024): tortured phrases mismatching their abbreviations (e.g., 'World Exchange Association (WTO)', the tortured version of 'World Trade Organization'). This approach is part of the Tortured Phrases ToolKit (TPTK) initiative, following the above presented research quality insurance workflow. To do so, we extracted all the abbreviations contained in the ELSST (n = 60) and THESOZ (n = 75) thesauri, we spun them using SpinBot, then we filtered out the unaltered abbreviations, ending with a total of 121 tortured abbreviations (Clausse *et al.*, 2025).

We included these generated tortured abbreviations in the PPS fingerprints list, then screened the scientific literature. We cross-validated these outputs using Dimensions, to filter out STEM-only documents according to the Australian and New Zealand Standard Research Classification (ANZRC) 2020 standards. We excluded paywalled articles, erroneous landing pages, defunct DOIs, and contents not featuring any tortured abbreviation.

We tested our approach on two case studies to screen articles included in the Hindawi Education Research International (EDRI) journal, and documents indexed by the GESIS Social Science Open Access Repository (SSOAR). We chose to do so as the open access publisher Hindawi, which is now owned by Wiley, has been the victim of a large-scale manipulation, leading to the publication of many paper-mill articles featuring tortured phrases. They released a full XML dump of their publications, which is unfortunately unavailable since June 2024. The EDRI journal contains a total of 760 articles. We also explored the GESIS SSOAR repository to ensure that no fraudulent content have been yet indexed. As of today, it contains 87,233 documents.

Thus, we used both the 42 tortured phrases and 121 tortured abbreviations (Clausse *et al.*, 2025) to explore the documents indexed by Dimensions, contained in the Hindawi EDRI journal, and indexed by the GESIS SSOAR (see Table 1). We lately extended the last exploration by mining all the abbreviations contained in the documents indexed in the latter repository, to find potentially new tortured abbreviations, and evaluate the generalization of the TPTK software.

Table 1. Examples of tortured phrases and tortured abbreviations related to HSS, used as fingerprints to flag problematic articles.

Tortured phrase	*Expected term*
Academic substantive information (PCK)	Pedagogical content knowledge (PCK)
Non-administrative associations (NGOs)	Non-governmental organizations (NGOs)
Communities for infectious prevention and anticipation (CDC)	Centers for disease control and prevention (CDC)
Uprightness of the votes	Electoral integrity
Trickery in conduct	Fraud
Geological locale	Geographical locations

Results and discussion

Exploring the PPS, we matched 543 documents featuring at least one of the 121 generated tortured abbreviations, and 107 additional documents matching at least one of the already referenced 42 HSS fingerprints. After filtering out the irrelevant articles and assessing the remaining ones, we found a total of **26 problematic documents**. We found between 1 and 6 distinct fingerprints in each document, these have been published between 2017 and 2024. They are either preprints from the Social Science Research Network (SSRN), then articles and proceedings from both local institutions and the 'haute couture' of scientific literature (such as Elsevier, IEEE, and Wiley).

Exploring the Hindawi EDRI journal, we found one article published in 2021, featuring 5 tortured abbreviations. Some of them were not yet part of the PPS fingerprint list. Following the same approach, we did not find any problematic document in the GESIS SSOAR repository. Finally, finding all the possible abbreviations in the GESIS SSOAR documents yielded a total of 23,477 abbreviations including 9,322 labelled as 'tortured', as depicted in Table 2.

Table 2. Processed documents from the GESIS SSOAR repository.

Type	*Count*
Total documents	87,233
English documents	33,748
Documents featuring abbreviations	23,477
Documents featuring tortured abbreviations	9,322
Validated false positives	5,048

We manually validated these results (as of today, we checked 5,048 of them), and the majority of them were false positives, such as foreign institutions (e.g., 'National Centre for Scientific Research (CNRS)' for the 'Centre National de la Recherche Scientifique') and reversed words (e.g., 'Hypothesis of Rational Expectations (REH)' instead of 'Rational Expectation Hypothesis'), making them genuine. However, we found 5 more problematic documents to be (re)assessed. These are preprints, articles, and monographs published between 2023 and 2024.

These 32 flagged documents are multidisciplinary and related to Education, Psychology, and Economics. Their screening highlighted new filtering rules to be implemented through the TPTK tortured abbreviations detector. As a contribution, we made new comments on the PubPeer platform to (re)assess 4 documents[1], which contain at least 4 distinct tortured abbreviations. We are aware that some of the matched abbreviations may still be false positives as they may have different meanings given the HSS field of research, and since they are less normalized as for STEM studies.

Conclusion

In this study, we explored the presence of tortured phrases in HSS articles. Using SpinBot, we generated 121 tortured abbreviations to be included in the PPS fingerprints list, in addition to the 42 HSS tortured phrases already referenced. We flagged a total of 32 multidisciplinary documents featuring tortured abbreviations related to Education, Psychology, and Economics, however we could not process the closed access ones since they are behind paywalls. We also found new filtering rules to be implemented through TPTK, to improve the precision of this software.

So far, we made 4 new comments on the PubPeer platform to alert readers that tortured phrases are also features in HSS articles. These flagged publications should be investigated by social scientists, as the domain experts working on the PPS are mostly related to STEM, and the HSS vocabulary is less normalized. However, more

[1] https://pubpeer.com/search?q=%22several+tortured+abbreviations%22

than just being aware of inconsistencies, actions should be taken against the articles assessed as fraudulent, by retracting them.
Finally, we proposed guidelines to encourage the scientific community to be aware of such fraudulent content, as a research quality insurance. We invite anyone interested in reassessing fraudulent articles to take part of the decontamination of the scientific literature, as an opportunity to embrace an established method.

Acknowledgments

We thank the political scientist Guillaume Levrier, who has been flagging several tortured phrases[2] in social sciences articles we could use in our study. We also would like to thank all the people who devote time and energy to decontaminate the scientific literature, mostly *pro bono*.
This article has benefited from an invited research stay at GESIS – Leibniz Institute for the Social Sciences, Köln, Germany via the Junior Research Call JRC-2024-02.
Alexandre Clausse and Guillaume Cabanac acknowledge the NanoBubbles project, that has received Synergy grant funding from the European Research Council (ERC), as part of the European Union's Horizon 2020 program, grant agreement number 951393. Fidan Badalova and Philipp Mayr received funding support from Deutsche Forschungsgemeinschaft (DFG), grant number MA 3964/15-3.

References

Abalkina, A., Aquarius, R., Bik, E., Bimler, D., Bishop, D., Byrne, J., Cabanac, G., Day, A., Labbé, C. & Wise, N. (2025). 'Stamp out paper mills' – science sleuths on how to fight fake research. *Nature*, 637(8048) (pp. 1047-1050). DOI: https://doi.org/10.1038/d41586-025-00212-1

Barbour, B. & Stell, B.-M. (2020). PubPeer: scientific assessment without metrics. In Biagioli, M. & Lippman, A. (Eds.), *Gaming the Metrics: Misconduct and Manipulation in Academic Research* (pp. 149-155). The MIT Press. DOI: https://doi.org/10.7551/mitpress/11087.003.0015

Biagioli, M. & Lippman, A. (Eds.). (2020). *Gaming the Metrics: Misconduct and Manipulation in Academic Research*. The MIT Press. DOI: https://doi.org/10.7551/mitpress/11087.001.0001

Cabanac, G. (2022). Decontamination of the scientific literature. arXiv preprint: https://arxiv.org/abs/2210.15912

Cabanac, G. (2024). Chain retraction: how to stop bad science propagating through the literature. *Nature*, 632(8027) (pp. 977-979). DOI: https://doi.org/10.1038/d41586-024-02747-1

Cabanac, G., Labbé, C. & Magazinov, A. (2021). Tortured phrases: a dubious writing style emerging in science. Evidence of Critical issues affecting established journals. arXiv preprint: https://arxiv.org/abs/2107.06751

Cabanac, G, Labbé, C. & Magazinov, A. (2022). The 'Problematic Paper Screener' automatically selects suspect publications for post-publication (re)assessment. arXiv preprint: https://arxiv.org/abs/2210.04895

Clausse, A., Cabanac, G., Cuxac, P. & Labbé, C. (2023). Mining tortured abbreviations in the scientific literature. *8th World Conference on Research Integrity* (WCRI'24). URL: https://hal.science/hal-04311600

[2] https://pubpeer.com/search?q=%22Guillaume+Levrier%22

Clausse, A., Badalova, F., Cabanac, G. & Mayr, P. (2025). Tortured Phrases from the Humanities and Social Sciences – A fingerprints dataset [Data set]. Zenodo. DOI: https://zenodo.org/records/14753785

Nazarovets, S. (2024). Dealing with research paper mills, tortured phrases, and data fabrication and falsification in scientific papers. In P.-B. Joshi, P.-P. Churi & M. Pandey (Eds.), *Scientific Publishing Ecosystem: An Author-Editor-Reviewer Axis* (pp. 233-254). Springer Nature. DOI: https://doi.org/10.1007/978-981-97-4060-4_14

O'Grady, C. (2024). Software that detects 'tortured acronyms' in research papers could help root out misconduct. *Science*. DOI: https://doi.org/10.1126/science.znqe1aq

Teixeira da Silva, J.-A. (2021). A tortured phrase claims heterosexuality of the carbon structure. *Results in Physics*, 30. DOI: https://doi.org/10.1016/j.rinp.2021.104842

Teixeira da Silva, J.-A. (2023). "Tortured phrases" in Covid-19 literature: can they serve as epistemic markers to assess the integrity of biomedical information? *Philosophy of Medicine*, 4(1). DOI: https://doi.org/10.5195/pom.2023.164

Van Noorden, R. (2023). How big is science's fake-paper problem? *Nature*, 623(7987) (pp. 466-467). DOI: https://doi.org/10.1038/d41586-023-03464-x

https://doi.org/10.51408/issi2025_116

What are the Most Important Elements of Research Activity to Assess? The Proposal of Relational Goods

Cinzia Daraio[1], Antonio Malo[2], Giulio Maspero[3], Ilaria Vigorelli[4]

[1]*daraio@diag.uniroma1.it*
DIAG Sapienza University of Rome, Via Ariosto, 25 00185 Rome (Italy)

[2]*malo@pusc.it*, [3]*maspero@pusc.it*, [4]*vigorelli@pusc.it*
Pontifical University of the Holy Cross, via dei Farnesi 83, 00186 Rome (Italy)

Abstract

The reform of research assessment has become a pressing concern for policymakers and institutions worldwide. In response to recent initiatives—most notably the European Commission's 2021 scoping report and the Agreement on Reforming Research Assessment—this paper offers a conceptual and practical contribution grounded in virtue ethics and relational sociology. We argue that to fully realise the aims of reform, research evaluation must move to include the *relational goods* produced within and between research practices. These goods—such as trust, collaboration, mentorship, and epistemic generosity—are essential for the sustainability and ethical integrity of scientific communities. Building on MacIntyre's theory of social practices and Donati's relational sociology, we propose a tripartite framework that integrates internal, external, and relational goods. We then outline a methodology for operationalising relational goods using qualitative and computational tools, including natural language processing and network analysis. By emphasising relationality as a criterion of research quality, this paper contributes to a paradigm shift in research assessment—one that is oriented toward social cohesion, virtue cultivation, and the flourishing of science as a human and communal endeavour.

Introduction

There is a growing interest in reviewing the methods used to evaluate research in Europe and beyond. The call for reform arises from a widespread recognition that current evaluation systems—largely dominated by publication metrics such as journal impact factors, citation counts, and university rankings—often fail to capture the richness, complexity, and societal value of research activity. Over time, such narrow metrics have shaped academic behaviour in unintended ways, promoting a culture of "publish or perish", undervaluing collaboration, diversity, and long-term societal impact, and limiting the visibility of contributions that do not align with mainstream academic norms.

In response to these concerns, the European Commission issued a scoping report in 2021 to lay the groundwork for rethinking research assessment in the European Research Area (ERA), stating that:

> "The proposed way forward [to reform current research evaluation systems] consists of a European agreement that would be signed by individual research funding organisations, research performing organisations and national/regional assessment authorities and agencies, as well as by their associations, all willing to reform the current research assessment system. The aim is for research and researchers to be evaluated based on their intrinsic

> merits and performance rather than on the number of publications and where these are published, promoting qualitative judgement with peer-review, supported by a more responsible use of quantitative indicators. The way in which the system is reformed should be appropriate for each type of assessment: research projects, researchers, research units, and research institutions. A reformed system should also be sufficiently flexible to accommodate the diversity of countries, disciplines, research cultures, research maturity levels, the specific missions of institutions, and career paths." (European Commission, 2021, p. 3).

This report emphasised the need to move beyond mechanistic and quantitative models of assessment and instead adopt qualitative, contextualised, and "responsible" approaches. It proposed the creation of a European agreement to be endorsed by a broad coalition of research actors—including funding organisations, research-performing institutions, and assessment bodies—willing to commit to reforming how research is evaluated across disciplines and contexts.

This proposal culminated in July 2022 with the release of the Agreement on Reforming Research Assessment, a milestone document outlining a shared vision and a set of ten core commitments to support systemic change. The ten main principles or *core commitments* are (see also Curry et al. 2020, which lists 15 manifestos reporting lists of principles for research assessment):

> "1. Recognise the diversity of contributions to, and careers in, research in accordance with the needs and nature of the research;
> 2. Base research assessment primarily on qualitative evaluation for which peer review is central, supported by responsible use of quantitative indicators;
> 3. Abandon inappropriate uses in research assessment of journal- and publication-based metrics, in particular inappropriate uses of Journal Impact Factor (JIF) and h-index;
> 4. Avoid the use of rankings of research organisations in research assessment;
> 5. Commit resources to reforming research assessment as is needed to achieve the organisational changes committed to;
> 6. Review and develop research assessment criteria, tools and processes;
> 7. Raise awareness of research assessment reform and provide transparent communication, guidance, and training on assessment criteria and processes as well as their use;
> 8. Exchange practices and experiences to enable mutual learning within and beyond the Coalition;
> 9. Communicate progress made on adherence to the Principles and implementation of the Commitments;
> 10. Evaluate practices, criteria and tools based on solid evidence and the state-of-the-art in research on research, and make data openly available for evidence gathering and research."

Among the most prominent principles are the need to: recognise the diversity of research outputs and careers; reduce reliance on journal- and publication-based metrics; center peer review in assessments; avoid inappropriate use of rankings; and

provide transparency, training, and accountability in the reform process. The Agreement has since gained significant traction and global resonance, giving rise to the Coalition for Advancing Research Assessment (CoARA, https://coara.eu/). The Coalition provides a platform for member organisations to collaborate, share practices, and collectively develop new tools and frameworks aligned with the Agreement's principles. The signatories of this Agreement agree on "the need to reform research assessment practices". Their shared vision is that:

> "the assessment of research, researchers and research organisations should recognise the diverse outputs, practices and activities that maximise the quality and impact of research. This requires basing assessment primarily on qualitative judgement, for which peer review is central, supported by responsible use of quantitative indicators. Among other purposes, this is fundamental for: deciding which researchers to recruit, promote or reward, selecting which research proposals to fund, and identifying which research units and organisations to support."

As of 15 April 2025, 774 organisations worldwide have joined CoARA, reflecting a strong and growing consensus across countries, institutions, and disciplines that research assessment must evolve to better serve science and society.

The shift toward more holistic and inclusive evaluation practices is not merely technical but fundamentally ethical and philosophical. It invites a rethinking of what counts as "good research," what values underpin scientific activity, and how excellence and impact are understood and rewarded. Our paper contributes to the broader reform movement by offering a novel conceptual lens: the centrality of *relational goods* within and across research practices.

We argue that research evaluation should move beyond focusing solely on tangible outputs—such as articles and patents—and instead recognise the social relationships, collaborative dynamics, and virtuous behaviours that sustain and enrich research as a human practice. Drawing from virtue ethics (MacIntyre, 1985) and relational sociology (Donati, 2010, 2019), we propose that assessing the quality of research should involve identifying and valuing the relational goods—such as trust, cooperation, mentorship, and epistemic generosity—that are essential for the flourishing of researchers, institutions, and the broader scientific community.

By integrating these philosophical and sociological perspectives, we seek to expand the normative foundations of research evaluation and to support the practical implementation of the CoARA principles. Our proposal invites stakeholders to see research not just as a competitive output-producing activity but as a cooperative and meaning-generating social endeavour, one that thrives through rich relational ecosystems.

Aim and contribution

This paper contributes to the ongoing reform of research assessment by proposing a conceptual and operational shift in how we understand and evaluate research activities. Our central thesis is that "relational goods"—the social, ethical, and cooperative dimensions that arise within and across research practices—represent the most significant, yet underappreciated, outputs of academic research.

Recognising and valuing these goods is critical for building an evaluation system that is not only technically robust but also ethically sound, socially responsive, and epistemically inclusive. We propose grounding research evaluation in a broader philosophical and sociological understanding of what constitutes "good research" and "good evaluation." Drawing on the virtue ethics of MacIntyre (1985) and the relational sociology of Donati (2010, 2019), we identify research practices as cooperative social endeavours whose excellence depends not only on technical outputs but also on the internal and external goods they generate, especially the relational ones.

Our contribution is threefold.

1) We extend existing frameworks by adding a third dimension to the established dual model of internal and external goods of research practices. We define relational goods as emergent, shared, and often intangible benefits—such as trust, mentorship, cooperation, and academic solidarity—that both sustain and transcend individual research practices. These goods are not reducible to material outputs or formal achievements, yet they are indispensable for long-term research vitality, epistemic integrity, and societal relevance.

2) By offering a rigorous ontological account of relational goods, we clarify their status as real and assessable elements of research ecosystems. We frame their evaluation within a normative perspective that privileges virtue ethics and the flourishing of researchers, enabling the design of assessment systems that prioritise human development, social cohesion, and epistemic justice.

3) We propose concrete tools for identifying relational dynamics within research outputs and communities. Our proposed framework provides evaluators and institutions with clear indicators and practices to incorporate relational quality into research assessments.

Through this integrative approach, the paper aims to bridge the gap between high-level policy declarations and the everyday realities of scientific work. It encourages institutions to design evaluation processes that value what makes research sustainable, collaborative, and socially embedded, thus contributing to a new culture of assessment grounded in relational excellence and virtue-oriented practice.

Materials and methods

In this paper, we contribute to the discussion on the reform of current research assessment practices by continuing and extending the analysis on "good evaluation" of research introduced in Daraio and Vaccari (2020, 2021 and 2022). In order to do a good evaluation, one must first know what good research consists of and use good research as the *normative* component of good evaluation. Daraio and Vaccari (2020) define a good evaluation as one that considers and emphasises good research. Good research was defined as that which takes place within the research practices considered as "social practice" according to MacIntyre (1985). A good evaluation of research practices, intended as social practices à la MacIntyre, should take into account the stable motivations and the traits of the characters (i.e. the virtues) of researchers.

This research line enables research to be assessed in the light of broad human interests and to take into account not only the outputs of research but also the psychology and motivation of researchers.
Specifically, Daraio and Vaccari (2020) use the notion of "good evaluation of research practices", characterising it as that evaluation that takes into account the constitutive elements of a "good research practice".
Following MacIntyre, Daraio and Vaccari (2020) propose to define a good social practice as

> "any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partially definitive of, that form of activity, with the result that human powers to achieve excellence, and human conceptions of the ends and goods involved, are systematically extended" (MacIntyre, 1985, p. 187).

Based on the definition of *good social practice*, they characterise a *good research practice* as

> "any coherent and complex form of socially established cooperative human activity through which its participants, through the exercise of a set of refined human psychological qualities or virtues, contribute to the advancement of the body of knowledge that is constitutive of that practice in a way that has a positive impact on the lives of researchers and society as a whole".

The most important elements of a good research practice (Daraio and Vaccari, 2020) are: i) internal and external goods and ii) the virtues of researchers. *Internal goods* of the practice are "high quality outcomes" of the practice that (a) can only be specified in terms of some specific practice (e.g. the way of conducting an empirical experiment; the practice of university teaching through lessons; the practice of interpretation of the text of classical authors in the humanities; etc.) and (b) can only be identified and recognized by the experience of participating in the practice in question. Those who lack the relevant experience are incompetent as judges of internal goods (MacIntyre 1985, p. 189);
(c) are typically achieved by those who follow the practice as an end in itself and enjoy the activities related to the practice;
(d) are typically achieved by those who experience gratitude towards teachers and mentors and justified anger towards those who betray our trust and violate our intellectual property;
(e) are typically achieved in conditions where one's potential and development are not hindered by fear and anxiety.
According to MacIntyre, internal goods include three kinds of outcomes: i) the high quality in performance (e.g. ability to question a text; ability to ask relevant questions during an experiment; ability to motivate one's research group or students in class, etc.); ii) the high quality of the outcome itself (e.g. articles, books, research projects,

discoveries, etc.); iii) the great value that comes from occupying a certain professional role in a research practice which contributes to the flourishing of researchers.

External goods are quality outcomes that are (a) only "externally and contingently attached" to the practice by the accidents of social circumstance and typically includes prestige, status and money, i.e. there are always alternative ways for achieving such goods, and their achievement is never to be had "only" by engaging in some particular kind of practice (MacIntyre, 1985, p. 201). And (b) when achieved, they are always some individual's property – i.e. the more someone has of them, the less there is for other people. They are characteristically objects of competition in which there must be losers as well as winners. On the contrary, internal goods include the outcome of competition to excel, but also positive externalities. This means that their achievement is good for the whole community that participates in the practice.

There is a rich literature on the analysis of research groups or teams based on network techniques (see e.g. Wuchty et al. 2007; Wang and Barabási, 2021). However, as noted by Bezuidenhout (2017, p. 1):

> "there is little literature that broadens out the scope of this analysis to consider the multidimensional nature of these research relationships. In particular, little is said about how scientists mediate their social interactions with peers during daily laboratory research. Less, indeed, is said about the tradition of 'learning through example' that characterizes most in situ laboratory training. All of these relational activities are of critical importance in sustaining and perpetuating the practice of science. It therefore becomes important to ask how we understand these relational activities directed towards building and sustaining relationships in different loci for the primary purpose of strengthening the practice of research and sustaining the traditions of scientific research".

Bezuidenhout (2017, p. 1) proposes a virtue ethics approach to understand these relationships using MacIntyre. In another work, Bezuidenhout and Warne (2018) propose to follow a theological approach to analyze the participation to research practices using the notion of ''callings'':

> "Callings highlight the identification and examination of individual talents to determine fit occupations for specific persons. Framing science as a calling represents a novel view of research that places the talents and dispositions of individuals and their relationship to the community at the center of flourishing practices".

Good scientists should have an *intuitive feeling* for their discipline, but they should also have a significant *personal satisfaction* from their work. They identify a key distinction between good and bad researchers considering personal joy in— and ''fittingness'' of—scientific occupations.

In this paper, we use philosophical argumentation to extend the conceptual and ontological framework currently adopted in the research evaluation reform debate. The prevailing direction is to adopt lists of principles of what evaluation should look like, detached from what research activity is. We start by defining what good research practices are, relying on MacIntyre, identifying good evaluation as that which is capable of enhancing good research practices. After that, we add the conceptual apparatus of "relational goods", developed in the new relational sociology, to extend good research practices and good evaluations of research practices to the connections that are in place at meso and macro levels.

Relational goods are neither material things nor benefits, but they have an economic, social and political value, as well as a moral and educational value.

According to Donati (2019) relational goods are relationships at the interpersonal level to the well-being social welfare of an entire community (friendship, trust, cooperation, reciprocity, social virtues, social cohesion, forgiveness given and received, solidarity and peace, complex societal relationships, such as the working climate in organizations, the sense of security or insecurity in the area in which we live, the relationships between family and work).

The notion of relational good emerges when we realise that there are "other" goods that are neither available on the basis of private proprietary title, nor accessible to everyone indiscriminately. They are goods that do not have an owner, nor are they of the collectivity generically understood. They are the goods of *human sociability*, goods crucial for the existence of society itself, which could not survive without them. If these goods are ignored, removed or repressed, the whole social fabric is impoverished, maimed, deprived of lifeblood, with serious damage to people and the overall social organisation. Relational goods (e.g., trust, cooperation, social virtues, and good working climate) are goods that offer the possibility of existence to the internal and external goods of research practices. In this sense, they exceed and encompass research practices by adding an important social dimension. This is why citations which are one of the most widely used indicators to measure the impact and quality of research but also to analyze collaborative networks between countries, authors or funding sources are not a relational good, but knowing the relational goods that produce the research practices in which the citations originate could be useful to qualify the nature of the citations, whether they originate from good or bad research practices, i.e., whether they are the result of self-citation networks that are self-sustaining in a publish or perish process or are the result of a genuine and wealthy knowledge creation process.

Preliminary Results and Discussion

A fundamental aspect to consider in reforming current evaluation practices is the *normative value* of good research practices for making a good evaluation of them. In this perspective, it is necessary to assess whether the practice of the academic/scientific research under examination is actually a good practice: (1) excellence of its outputs; (2) the way in which they are achieved (in accordance with the rules that constitute the practice); (3) the impact that following the practice has on researchers' life plans. But also, external goods should be taken into account.

Moreover, it is crucial to take into account researchers' virtues, i.e. stable traits of character that make it possible to grasp and pursue the internal goods of research practices. In order to take account of both internal and external goods, the evaluation of research practice must also be able to assess the ability of researchers to obtain them, i.e. the virtues of the participants in the practice. According to MacIntyre, virtue is

> "an acquired human quality the possession and exercise of which tends to enable us to achieve those goods which are internal to practices and the lack of which effectively prevents us from achieving any such goods (Macintyre, 1985, p. 191)".

In this paper, we consider research practices as the departure point from which relational goods are built and developed. Donati (2010) shows that relational goods have their ontological reality and are endowed with the following properties: (i) they consist of *social relations* that are not reducible to mere interactions or transactions (and therefore different from market goods); (ii) these social relational goods are an *emergent effect* with respect to the contributions made by the subjects in the relationship; (iii) as relations, these goods possess a *reality sui generis*, that is, they have a certain structure, which is processual and changes over time; (iv) they are produced and enjoyed together by those who participate in them; (v) they bring benefits both to the participants and to those who share their reflections from the outside, without that *none of the individual subjects can appropriate them alone*. These characteristics differentiate relational goods from public goods, market goods and externalities.

A relational good refers to the good found in "being in a (certain) relationship". It is therefore crucial to understand what "being in relation" means. This expression can be declined in two ways: either as "the fact of being in relation" or also as "the being that is (what there is) in relation". According to Donati (2010)'s point of view, "being in relationship" is an expression that has three analytical meanings: (i) the fact that between two (or more) entities there exists a certain *distance* which, at the same time, distinguishes and connects these entities; (ii) that this relation exists in the sense that it has its *own reality* with its own causal powers; (iii) that such reality has its *own mode of being* (the mode of being that is in the relation). This perspective of social ontology demands to be translated into a sociological discourse, which is moreover amenable to empirical research.

Relational goods are conditions of possibility of research practice from which the internal and external goods of research itself flow. For example, the "organizational climate" of a research group plays an important role in the possibility of the research group to achieve the internal goods of the practice, excelling in the same, and also to achieve external goods in order to sustain and develop the practice itself.

The whole research practice should be analyzed, including "other characteristics" that connect the practice to its broader relational social dimension. In the quantitative evaluation we should consider: - A "relational accountability" of public investment when assessing *research performing organisations and research units* for funding

allocation; - A "relational project management" that enhances future research funding decisions when assessing *research projects* for funding allocation promoting; - Valuing the "relational aspects" of intellectual virtues when assessing *individual researchers and research teams* for funding allocation, recruitment and hiring promotion, professional development review, prize and award decisions.

We will discuss how the development of a *relational virtue ethics* can contribute to the identification of more relevant aspects of research activity to be evaluated and valued. We will try to show that relational goods, as "latent goods", can be measured only indirectly, through observable proxies that might be found in the analysis of the virtues of researchers, groups and institutions that comprise them.

Finally, the characterization of the internal and external goods produced within the research practices taking into account the relational goods that generate the research practice will allow us to provide a *hierarchy* of the three missions of universities and research centers, that are teaching, research and the so called "third mission" (or knowledge transfer and impact on the society in general terms).

From Theoretical Argumentation to Operational Pathways: Towards a Concrete Evaluation of Relational Goods

In this section, we propose a concrete operationalisation of the concept of relational goods in research. This is intended to clarify how the abstract theoretical foundations of our proposal can be translated into tangible evaluative practices. In doing so, we aim to provide scholars, evaluators, and institutions with the tools to observe, interpret, and eventually assess the relational quality of research activity, both in its textual expression and in the wider social ecosystem of scientific collaboration.

The analysis of relational goods may be meaningfully approached through two complementary entry points. The first concerns the *internal structure of the scientific text*—how collaboration and cooperation manifest within the citation practices, authorship patterns, and narrative voice of the article itself. The second concerns the *broader external relations of the research activity*, such as inter-institutional collaboration, mentoring structures, and team governance. These two fronts—internal and external—reveal relational goods as they are embedded within and extend beyond individual research outputs.

Internally, we argue that relational goods can be discerned by analysing how previous literature is engaged. This requires more than counting citations; it necessitates an interpretive reading of the relational intent of each citation. A citation may be supportive, building upon a previous result and weaving it into a shared research lineage, or it may be oppositional, serving to challenge or distance the cited claim. While both are legitimate forms of scholarly engagement, their relational valence differs significantly. The cooperative quality of research is often higher when a work integrates and acknowledges the epistemic contributions of others in a generative and dialogical manner, as discussed in the virtue ethics approach proposed by Bezuidenhout (2017).

Another important signal lies in authorship patterns, particularly across generations. Co-authorships involving both senior and early-career researchers may reflect practices of mentoring and transmission of expertise, which we interpret as

expressions of magisteriality. Such forms of cooperation are central to the generativity of research groups and the sustainable reproduction of knowledge communities (Bezuidenhout & Warne, 2018). Likewise, ethical citation practices—such as acknowledging underrepresented voices or non-mainstream sources—may point to a virtue-oriented scholarly style, highlighting academic generosity and inclusiveness (MacIntyre, 1985; Daraio & Vaccari, 2020).

Externally, relational goods manifest in the enduring connections that research groups and institutions form with each other. Collaborative networks that are long-standing and rooted in mutual respect, rather than opportunistic partnerships, can be identified through bibliometric indicators such as the frequency and longevity of co-authorship between institutions. Cross-cultural and interdisciplinary collaborations, when grounded in shared intellectual aims, often reflect high levels of relational trust and openness (Wuchty et al., 2007; Wang & Barabási, 2021).

Mentoring networks, although often informal, can be traced through structured data on academic genealogy and project leadership. Furthermore, institutional practices—such as fair authorship distribution, shared leadership, and inclusive project design—serve as indicators of a virtuous research culture that nurtures relational goods. These practices have been discussed in recent sociologies of scientific collaboration, which emphasise the ethical and social conditions under which scientific excellence is pursued (Donati, 2019; Nowotny et al., 2001).

To make these dimensions empirically tractable, we propose the use of artificial intelligence and bibliometric tools to support analysis. Natural language processing (NLP) can be used to detect relational cues in citation contexts, distinguishing between citations that build upon, contrast with, or merely acknowledge prior work. Network analysis tools can map the structure and quality of co-authorship and collaboration patterns, revealing not only who collaborates but also how these collaborations evolve. Topic modelling can help identify the cohesion and epistemic continuity within research teams, while demographic inference methods can be used to detect generational patterns in authorship, pointing to possible mentoring dynamics.

This dual approach—attending both to the textual dimension of research and to its institutional-relational context—provides a structured pathway to identify, interpret, and assess relational goods.

To summarise the proposed operational framework, we provide below a summary table (Table 1) that organises the main components of our analysis. This table identifies key dimensions through which relational goods in research can be observed, the specific indicators relevant to each area, and the methodological tools that can support their assessment. It serves as a bridge between our theoretical arguments and their empirical implementation, illustrating how internal textual elements and external relational dynamics can be systematically analysed using qualitative and computational methods. The inclusion of AI-assisted tools highlights the feasibility of scaling this framework across diverse research contexts.

Table 1. Operationalizing Relational Goods in Research Evaluation.

Dimension	Focus Area	Indicators / Elements	Analytical Approach / Tools
Internal	*Citation Intent*	Supportive vs. Oppositional references; dialogical integration or critique	Qualitative citation context analysis (NLP)
	Generational Dialogue	Presence of intergenerational co-authorship; evidence of mentoring relationships	Co-authorship metadata; demographic inference
	Ethical Citation Practices	Inclusion of underrepresented authors or schools; epistemic generosity	Bibliographic diversity measures; citation context classification
External	*Collaboration Patterns*	Longevity and frequency of institutional collaborations; cross-cultural teams	Network analysis; co-authorship graphs
	Mentoring Networks	Academic genealogies, team continuity, senior-junior linkages	Project funding databases; ORCID data; CV parsing
	Virtuous Group Practices	Fair authorship ordering; inclusive decision-making; and leadership rotation	Institutional policies; team-level ethnographic study
Transversal AI tools	*AI-Supported Analysis*	Tools for identifying relational patterns from large-scale data	NLP (citation sentiment); network science; topic modelling

Conclusions

In this paper, we have engaged with the current movement toward reforming research assessment practices, offering both a theoretical deepening and a practical extension of the debate. While recent policy initiatives—such as the European Commission's scoping report and the Agreement on Reforming Research Assessment—mark an important shift in recognizing the limitations of metric-driven evaluations, we argue that a more fundamental rethinking is needed. This rethinking must begin by asking: what constitutes good research, and what does it mean to evaluate it well?

Our central contribution is the proposal to redefine the key outputs of research activity through the lens of relational goods. We suggest that alongside internal and external goods, relational goods—such as trust, collaboration, mentorship, epistemic

generosity, and social cohesion—are fundamental to the vitality of research communities and the broader scientific enterprise. These goods are not merely incidental to knowledge production; they are constitutive of research quality itself, supporting sustainable excellence, interdisciplinary dialogue, and the ethical formation of researchers.

Grounding our analysis in MacIntyre's virtue ethics and Donati's relational sociology, we have outlined a normative and ontological framework that highlights the ethical dimensions of research practice. In doing so, we have positioned research not only as a technical or productive activity but as a moral and relational practice, embedded in networks of cooperation, mentoring, and shared inquiry.

Importantly, we have translated this conceptual apparatus into a concrete operational framework, offering institutions, evaluators, and policymakers a practical path forward. Through internal indicators (e.g., citation intent, ethical citation practices, intergenerational co-authorship) and external indicators (e.g., collaborative networks, mentoring structures, team governance), we propose a multi-layered methodology that can help identify and assess the presence and quality of relational goods within research ecosystems. The integration of artificial intelligence tools, such as natural language processing and network analysis, further enhances the feasibility and scalability of this approach.

By embracing relational goods as core evaluative dimensions, we propose a shift from output-centred assessment to a relationally-anchored evaluation paradigm—one that emphasises sustainability, inclusion, and the long-term flourishing of researchers, institutions, and society at large. This perspective not only aligns with the principles promoted by CoARA and similar reform movements but deepens their foundations by offering a clear philosophical justification and an actionable roadmap.

Ultimately, we envision a model of research evaluation that values the virtue-driven and socially embedded nature of research, recognising excellence not only in individual achievements but also in the quality of relationships, the strength of collaborative cultures, and the generativity of academic communities. Such a model can enable a more just, reflective, and human-centred scientific enterprise—one in which evaluation serves to enhance rather than constrain the deeper purposes of research.

References

Bezuidenhout, L. (2017). The relational responsibilities of scientists: (Re) considering science as a practice. *Research Ethics*, 13(2), 65-83.

Bezuidenhout, L., Warne, N. A. (2018). Should we all be scientists? Re-thinking laboratory research as a calling. *Science and Engineering Ethics*, 24(4), 1161-1179.

Curry, S., de Rijcke, S., Hatch, A. et al. (2020), The changing role of funders in responsible research assessment: progress, obstacles and the way ahead. Working Paper. Research on Research Institute (RoRI) https://doi.org/10.6084/m9.figshare.13227914.v1

Daraio, C., Vaccari, A. (2020). Using normative ethics for building a good evaluation of research practices: towards the assessment of researcher's virtues. *Scientometrics*, 125(2), 1053-1075.

Daraio C. Vaccari A. (2021), Perché è importante fare una buona valutazione della ricerca. La proposta delle virtù, Bollettino della Società Filosofica Italiana, gennaio-aprile 2021, pp. 45-59.
Daraio C. Vaccari A. (2022), How should evaluation be? Is a good evaluation of research also just? Towards the implementation of good evaluation, *Scientometrics*, DOI 10.1007/s11192-022-04329-2.
Donati, P. (2010). *Relational sociology: A new paradigm for the social sciences*. Routledge.
Donati, P., Archer, M. S. (2015). *The relational subject*. Cambridge University Press.
Donati P. (2019), *Scoprire i beni relazionali*, Rubbettino.
Donati, P. (2021). *Transcending modernity with relational thinking*. Taylor & Francis.
European Commission (2021), Towards a reform of the research assessment system: scoping report. November 2021, Bruxelles, ISBN 978-92-76-43463-4.
MacIntyre, A. (1985). *After virtue*. Duckworth.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.
Nowotny, H., Scott, P., Gibbons, M. (2001). *Re-thinking science: Knowledge and the public in an age of uncertainty*. Cambridge: Polity.
Wang, D., Barabási, A. L. (2021). *The science of science*. Cambridge University Press.
Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

https://doi.org/10.51408/issi2025_102

Will Scientific Research Drive Technology to be a Hit? A Comparison between Emerging Technological Fields and Traditional Technological Fields

Xi Chen[1], Jin Mao[2], Gang Li[3], Xuehua Wu[4]

[1] sjcnh9956@whu.edu.cn, [2] danveno@163.com, [3] imiswhu@aliyun.com, [4] uxuehua@whu.edu.cn
Wuhan University, Center for Studies of Information Resources, Wuhan (China)
Wuhan University, School of Information Management, Wuhan (China)

Abstract

Translating scientific knowledge into viable technologies demands specialized efforts. The Linear Model, an early conceptual framework for understanding this process, is widely used in science-intensive sectors. Patent citations to scientific literature often measure the reliance of technology on science, but most studies focus on document-level analysis. However, they may fail to capture the full scope of the development and interconnections of technologies. This study identified the take-off times of technology trajectories and distinguished emerging technological fields (ETFs) from traditional technological fields (TTFs). We measured the distance of each field from the "paper-patent boundary" and conducted a comparative analysis between ETFs and TTFs. Additionally, we defined and calculated scientific connectivity within these fields to evaluate their integration of technology and science. Our findings show that ETFs experience more significant fluctuations in their distance to the paper-patent boundary over time and consistently exhibit higher scientific connectivity despite the divergence from the academic frontier. This study advances the understanding of knowledge transfer from science to technology, offering valuable insights in how scientific research fosters innovation.

Introduction

Scientific research forms the cornerstone of novel inventions, generating a wealth of valuable ideas that drive technological progress (Fleming & Sorenson, 2004; Chen, Mao, & Li, 2024). Since Narin and Olivastro's (1992) seminal work, growing evidence has shown that patent citations to scientific literature indicate knowledge transfer from science to technology. Most studies analyze this transfer at the document level, focusing on how discoveries from scientific publications lead to new inventions in specific fields. However, technologies rarely rely on a single invention; instead, they evolve through a developmental process, producing successive inventions that refine or expand their applications (Arthur, 2007). This progression allows technologies to increase their impact over time. Evolutionary economists have framed this structured development as progress along established trajectories (Dosi, 1982). To fully understand how scientific research drives technological progress, it is essential to examine its role in shaping technology trajectories rather than focusing solely on individual inventions.

The Linear Model outlines a progression from basic research to applied research, followed by development, production, and, finally, diffusion (Balconi et al., 2010; Bush, 2021). Although this model has faced criticism for implying that basic research is not always directly linked to technological progress, Balconi et al. (2010) highlight the critical role of knowledge supply in fostering industry development in science-

intensive sectors. Nonetheless, it remains unclear whether emerging technological fields maintain a closer relationship with scientific research compared to traditional technological fields.

In this study, we identified the Take-off time of technology trajectories and accordingly distinguished ETFs from TTFs. We subsequently measured the distance of each technological field to the "paper-patent boundary" and conducted a field-level comparison between ETFs and TTFs. This distance quantifies the proximity of a technological field to scientific research (Ahmadpoor & Jones, 2017). It essentially captures the translation path from scientific discoveries to technological innovations. Then we defined the scientific connectivity, which assesses the overall integration of technology with science within a given field. It reflects how much patents within a field operate in independent or overlapping fields relative to scientific work. Lastly, we examined the relationship between the distance and the scientific connectivity.

This ongoing study aims to reveal whether scientific research can drive technology to take off, or, in other words, be a hit. It highlights how ETFs and TTFs evolve in their reliance on scientific research along their technology trajectories.

Method

Data

This study collected utility patents from the USPTO and analyzed them at the patent family level to account for similar technical subject matter across different inventions. To account for the time lag between filing and granting, we limited the filing year to 2014. The final sample includes 3,105,854 patents filed between 1976 and 2014, belonging to 2,469,053 patent families.

Scientific references in patents are obtained from Reliance on Science (Marx and Fuegi, 2020a, 2020b), which contains 40,393,301 citations to papers from patents. We collected 2,728,680 paper-patent citations in 474,633 patent families between 1976 and 2014 using citations with a confidence score of 10. Each patent family cites an average of 5.75 scientific references.

Identification of emerging technological fields

We mapped the trends of technological fields by tracking the cumulative number of patent families filed and granted per year, using 4-digit IPC codes. To differentiate ETFs from TTFs, we identified the trajectory's *Take-off time* and *Technological impact* based on Pezzoni et al. (2022)'s method. *Technological impact* is measured by the cumulative number of patents a technology accumulates over 20 years. A technology reaches "takeoff" when it attains a specific percentage of its maximum technological impact (Griliches, 1957; Pezzoni et al., 2022). *Take-off time* refers to the number of years that pass from the appearance of a technological field until its contribution to the *Technological impact* reaches 10% (Pezzoni et al., 2022).

To mitigate the risk of underestimating the maximum technological impact of the technologies with a late takeoff, we fitted a trend function using the observed cumulated distribution of subsequent patent families.

$$Num_t = \frac{Ceiling}{1+e^{(\frac{t-Midpoint}{\alpha})}} \text{ (Pezzoni et al., 2022)} \quad (1)$$

where Num_t is the cumulative number of patent families predicted at year t; t is the number of years that pass from the appearance of a technological field; the parameter *Ceiling* is defined as the upper asymptote of the S-curve; *Midpoint* is the required time to reach 50% of the ceiling; α is the inverse of the curve slope at the *Midpoint*. The estimated take-off time can be determined by linearly combining the predicted trajectory's *Midpoint* and α parameters: *Take off* = *Midpoint*-2.2*α.

Identification of reliance on science

Distance to the "paper-patent boundary"

To assess the extent to which technological fields depend on science, we used the concepts of the "paper-patent boundary" and "distance to the boundary" (Ahmadpoor & Jones, 2017). The "paper-patent boundary" represents direct patent citations to academic papers within an integrated citation network. We then calculated the minimum citation distance of all other patents from this boundary. This approach maps the interface between scientific research and technological innovation, illustrating how discoveries transition into applications.

The distance to the "paper-patent boundary" was denoted as D_i for each patent i. When a patent directly cites a paper, $D_i = 1$, representing the patent is at the "paper-patent boundary". For the other patents, a patent i with $D_i = n+1$ is one that cites a patent j with $D_j = n$ and does not cite any patent k with $D_k < n$. Patents that are incapable of being linked at any distance to the "paper-patent boundary" are characterized as "unconnected." The process is shown in Figure 1(A). Subsequently, the distance to the "paper-patent boundary" was quantified by averaging the values of D_i within a patent family, thereby assessing its reliance on papers. Figure 1(B) illustrates that about 78% patents can be traced to scientific research.

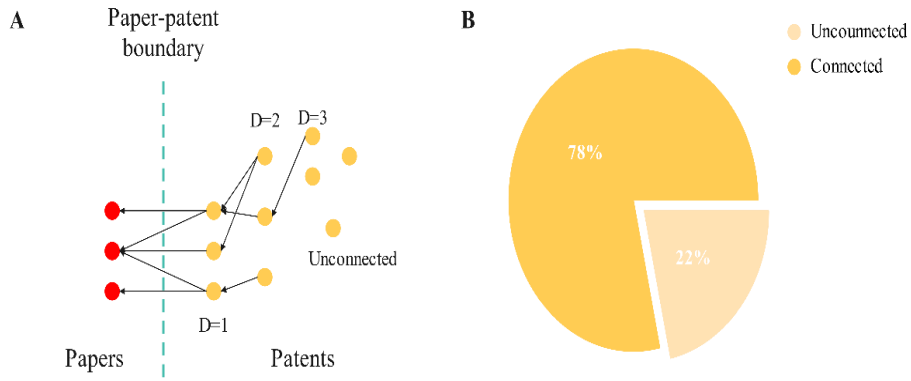


Figure 1. (A) The integrated citation network from patents to papers and the distance to the "paper-patent boundary". (B) The proportion of patents with backward links to a paper.

Scientific connectivity

Scientific connectivity reflects the extent to which patents exist in independent spheres, serving as a measure of the overall integration of technology with scientific research. It is:

$$SC = \frac{PT'}{PT} \tag{2}$$

where SC is scientific connectivity of a technological field; PT represents the number of patent families in a field; PT' is the number of patent families can be traced to papers.

Results

Distance to the "paper-patent boundary"

In this section, we evaluated the extent to which ETFs and TTFs rely on scientific research by comparing their distance to the "paper-patent boundary."

First, we categorized technological fields as either emerging or traditional. As shown in Figure 2(A), technological fields that took off after 2004 were labelled as ETFs and there are 191 ETFs and 455 TTFs. Figure 2(B) illustrates the mean distance to the paper-patent boundary for all patent families in a technological field. The plot shows that figures for both TTFs and ETFs range from 2.3 to 3.4.

Figure 3 shows the distribution of the mean distance to the paper-patent boundary of a technological field according to the take-off time. ETFs primarily concentrate their take-off time within the ranges of 1–12 years and 25–35 years. ETFs with different take-off times demonstrate a higher degree of fluctuation in their distance to the paper-patent boundary. In comparison, regardless of the take-off time, TTFs exhibit relatively stable distance to the paper-patent boundary. The interquartile ranges across the box plots generally remain consistent, indicating less variability in how they connect with scientific research.

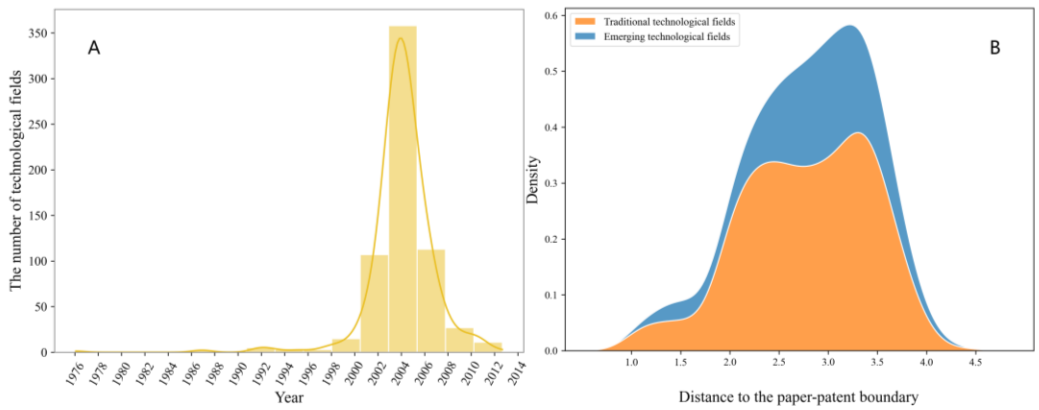


Figure 2. (A) The distribution of take-off times across all technological fields. (B) The distribution of the mean distance to the paper-patent boundary in a technological field.

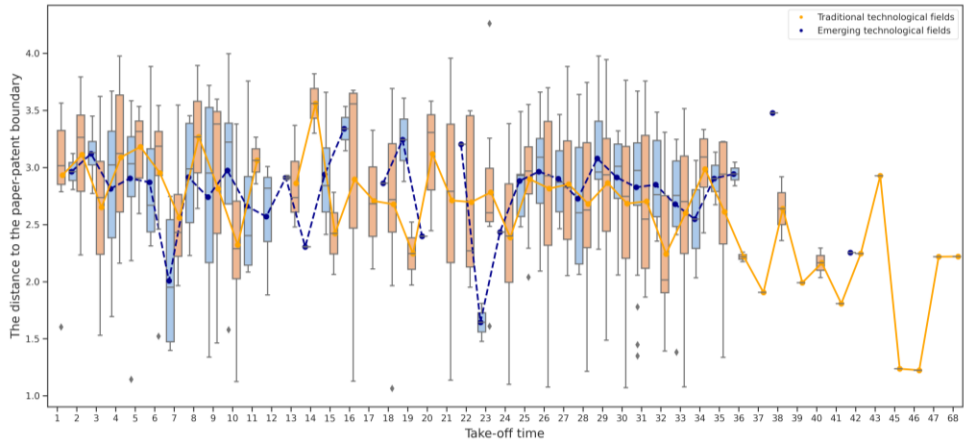


Figure 3. The distribution of the mean distance to the paper-patent boundary of a technological field according to the take-off time.

Figure 4 illustrates how the mean distance to the paper-patent boundary changes with takeoff status. ETFs consistently maintain a greater mean distance than TTFs. Before takeoff, both have a low mean distance, indicating strong ties to scientific research. After takeoff, the mean distance increases for both, but ETFs diverge more rapidly, suggesting a quicker shift toward practical applications. TTFs increase their distance more gradually. In later stages, TTFs slightly decrease their distance, indicating realignment with academic research, while ETFs also exhibit a gradual reduction, suggesting a renewed connection to science over time.

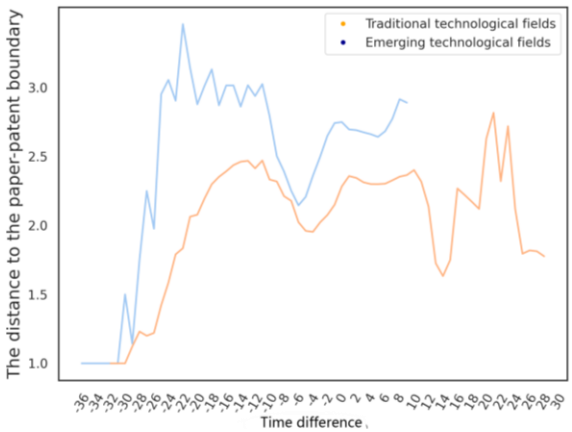


Figure 4. The mean distance to the paper-patent boundary of TTFs and ETFs. The x-axis represents a time difference metric, where the value is obtained by subtracting the natural year from the take-off year. Negative values indicate years before the technology reached its 10% impact threshold (i.e., pre-takeoff), while positive values indicate years after the technology achieved its takeoff.

Scientific connectivity

This section examines the relationship between the proximity of a technological field to scientific research and its integration with science by analysing scientific connectivity in TTFs and ETFs at varying distances from the paper-patent boundary. Figure 5(A) shows that ETFs exhibit decreasing scientific connectivity as they move away from the boundary, with high connectivity and low variability at the closest distance (D=1). In contrast, TTFs show increasing scientific connectivity with distance, starting lower than ETFs at D=1 but becoming more connected over time. This suggests that ETFs tend to cite patents derived from academic papers. Despite distance, ETFs consistently maintain higher scientific connectivity than TTFs. Figure 5(B) illustrates the mean scientific connectivity of TTFs and ETFs across different takeoff states. Both exhibit similar trends over time, with fluctuations before takeoff and convergence toward stability in the post-takeoff period.

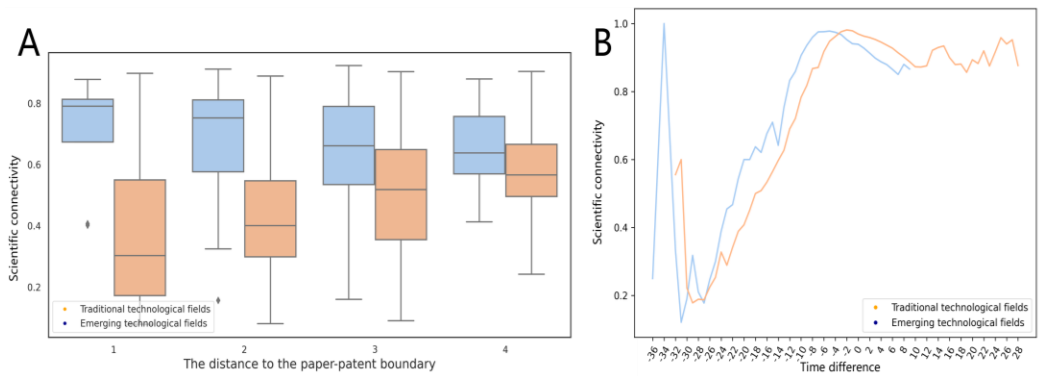


Figure 5. (A)The distribution of scientific connectivity at different distances to the paper-patent boundary. (B)The mean scientific connectivity of TTFs and ETFs.

Discussion and conclusion

This study investigates whether scientific research can serve as a catalyst for the takeoff of technologies by examining how ETFs and TTFs evolve in their reliance on it. The results show:

First, ETFs experience greater fluctuations in their distance from the paper-patent boundary over time, while TTFs follow a more stable trajectory after taking off. This suggests that emerging technologies, initially driven by scientific research, rapidly shift toward practical applications, temporarily diverging from academia (Stahl et al., 2017). However, as these fields mature, they realign with scientific research, possibly due to the convergence of academic advancements with practical needs or new research emerging in response to industry demands.

Second, the declining scientific connectivity in ETFs as they move away from the paper-patent boundary indicates that these fields start with strong academic foundations but gradually transition toward commercialization (Islam et al., 2018). In contrast, the increasing connectivity in TTFs suggests a cyclical relationship with research—initially shifting away from academia to refine and apply existing

knowledge but later returning to academic research to address new challenges and drive further innovation.

Additionally, the consistently higher scientific connectivity in ETFs, even as they move away from the academic frontier, highlights the critical role of scientific research in emerging technologies. This underscores the need for ongoing collaboration between academia and industry to sustain innovation.

This study enhances the understanding of knowledge transfer from science to technology, offering insights into how scientific research shapes technological trajectories. It also clarifies when and how fields transition from research-driven innovation to application-focused development.

A key limitation of this study is the reliance on patent citations as a measure of the science-technology relationship. While useful, this metric may not fully capture the complexity of knowledge transfer. Future research should explore alternative indicators to provide a more comprehensive view of how science influences technological advancement.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (NSFC), Grant Nos. 71921002 and 72174154.

References

Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. Science, 357(6351), 583-587. https://doi.org/10.1126/science.aam9527

Arthur, W. B. (2007). The structure of invention. Research Policy, 36(2), 274-287.

Balconi, M., Brusoni, S., & Orsenigo, L. (2010). In defence of the linear model: An essay. Research Policy, 39(1), 1-13. https://doi.org/10.1016/j.respol.2009.09.013

Bush, V. (2021). Science, the endless frontier. http://digital.casalini.it/9780691201658

Chen, X., Mao, J., & Li, G. (2024). A co-citation approach to the analysis on the interaction between scientific and technological knowledge. Journal of Informetrics, 18(3), 101548. https://doi.org/10.1016/j.joi.2024.101548

Chen, X., Mao, J., Ma, Y., & Li, G. (2024). The knowledge linkage between science and technology influences corporate technological innovation: Evidence from scientific publications and patents. Technological Forecasting and Social Change, 198, 122985. https://doi.org/10.1016/j.techfore.2023.122985

Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. Research Policy, 11(3), 147-162.

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. Strategic Management Journal, 25(89), 909-928. https://doi.org/10.1002/smj.384

Griliches, Z. (1957). Hybrid corn: An exploration in economics of technological change (Doctoral dissertation, The University of Chicago). https://www.proquest.com/dissertations-theses/hybrid-corn-exploration-economics-technological/docview/301928266/se-2?accountid=9652

Islam, M., Fremeth, A., & Marcus, A. (2018). Signaling by early-stage startups: US government research grants and venture capital funding. Journal of Business Venturing, 33(1), 35-51. https://doi.org/10.1016/j.jbusvent.2017.10.001

Marx, M., & Fuegi, A. (2020a). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. Journal of Economics & Management Strategy, 31(2), 369-392. https://doi.org/10.1111/jems.12455

Marx, M., & Fuegi, A. (2020b). Reliance on science: Worldwide front-page patent citations to scientific articles. Strategic Management Journal, 41(9), 1572-1594. https://doi.org/10.1002/smj.3145

Narin, F., & Olivastro, D. (1992). Status report: Linkage between technology and science. Research Policy, 21(3), 237-249.

Pezzoni, M., Veugelers, R., & Visentin, F. (2022). How fast is this novel technology going to be a hit? Antecedents predicting follow-on inventions. Research Policy, 51(3).

Stahl, B. C., Timmermans, J., & Flick, C. (2017). Ethics of emerging information and communication technologies: On the implementation of responsible research and innovation. Science and Public Policy, 44(3), 369-381. https://doi.org/10.1093/scipol/scw069

Wang, J. J., & Fred, Y. Y. (2021). Probing into the interactions between papers and patents of new CRISPR/CAS9 technology: A citation comparison. Journal of Informetrics, 15(4), 101189. https://doi.org/10.1016/j.joi.2021.101189

POSTER

https://doi.org/10.51408/issi2025_196

15 Years of the Eastern Partnership Initiative: A Bibliometric Reflection

Maria Ohanyan[1], Mariam Yeghikyan[2], Miranush Kesoyan[3], Simon Hunanyan[4], Aram Mirzoyan[5]

[1]mohanyan226@gmail.com, [2]mariam_yeghikian@mail.ru, [3]mkesoyan1996@gmail.com, [4]simhunanyan@gmail.com, [5]aram.mirzoyan@asnet.am
Institute for Informatics and Automation Problems of NAS RA, Center for Scientific Information Analysis and Monitoring, 1, P. Sevak str., 0014 Yerevan (Armenia)

Introduction

After the collapse of the Soviet Union, the European Union (EU) played a key role in promoting regional cooperation among former Soviet states. The EU focused on strengthening relations both with these post-Soviet countries and among them (Delcour, 2011). The EU extended its borders to several Eastern countries with historically weaker economic ties. These countries are characterized by significant institutional and structural differences. The European Union's regional cooperation policy framework aimed at enhancing prosperity, stability, and security, to create a 'ring of friends' and extend the EU's influence to its neighboring regions (Petrakos et al., 2015). This initiative led to the creation of the Eastern Partnership (EaP) on May 7, 2009, which is a specific Eastern dimension of the European Neighbourhood Policy. The creation of the EaP was part of the EU's broader strategy to strengthen ties with six Eastern European countries: Armenia, Azerbaijan, Belarus, Georgia, Moldova, and Ukraine (Korosteleva, 2011). The initiative aims to strengthen and deepen political, economic, and scientific relations between the EU, its Member States, and partner countries while supporting sustainable reform processes across the Eastern Partnership region (Eastern Partnership, 2009).

The European Union's integration policy has spurred extensive scholarly research, particularly focusing on the EU's pivotal role as a global actor (Delcour, 2011). Academic discussions have increasingly emphasized the EU's engagement with its Eastern neighbors through the Eastern Partnership initiative.

2024 marks the 15th anniversary of the Eastern Partnership initiative. It is an opportune moment to reflect on the journey taken and evaluate the results achieved. There are various ways to do this, including bibliometric analysis. In this poster, we present the 15-year journey of the EaP based on the data collected from Web of Science. We aim to analyse the distribution of publication dynamics over these years, the involvement of the six-member countries in academic publications, and the relevant academic fields represented in these works. This will enable to picture of the bibliometric reflection of EaP since 2009.

Data and methodology

The data was collected from the Web of Science (WOS) international scientific database for the years 2009–2024. The search was conducted using the terms 'Eastern Partnership' and 'EaP' across all fields, returning 353 items. After data cleaning, the final result was 317 items, which refer directly to the EaP. The data was processed and illustrated in graphs to highlight publication dynamics over the years and the distribution across the six EaP countries and EU member states. The academic fields and disciplines of the publications were categorized according to Glänzel and Schubert's classification (Glänzel & Schubert, 2003). We used the full counting method, assigning full value to publications that covered all specified areas and countries simultaneously.

Results

The study found that 317 publications on the Eastern Partnership (EaP) were indexed in the WOS database between 2009 and 2024. The top 3 document types are Articles (249),

Book Chapters (70), and Conference Proceedings (48). The highest number of publications was recorded in 2014 and 2016. In the following years, the number of publications declined. However, since 2021, the number of publications has steadily increased and continues to rise through 2024 (Fig. 1).



Figure 1. Number of publications in WOS on EaP from 2009 to 2024.

When analysing the distribution of publications by country, we divided them into two main groups: EU member countries (Fig. 2) and EaP countries (Fig. 3). For the former, we observe the following pattern:

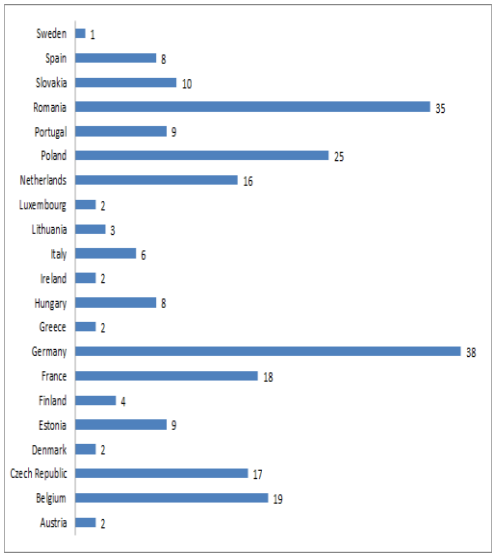


Figure 2. The number of articles involving researchers from EU countries.

And for the latter, there is the following picture:

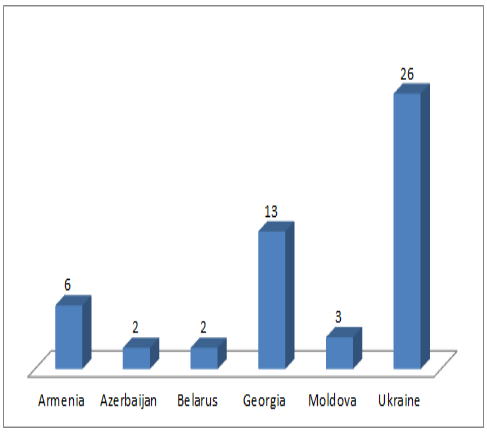


Figure 3. The number of articles involving researchers from EaP countries.

It is important to note that since June 2021, Belarus has suspended its participation in the EaP. However, the EU continues to maintain cooperation with Belarusian civil society.

When turning to the countries that are neither from the EU nor from EaP, the top 3 countries with the most publications are the UK[1] (35), Russia (19), and the US (15).

The results shown in Fig. 4 indicate that publications have been distributed across various fields as follows: the highest number of publications belongs to Social Science II (226) and Social Science I (119). The next two most popular fields—Arts and Humanities (10) and Multidisciplinary Sciences (8)—are far behind. Other fields have even fewer cases or none at all:

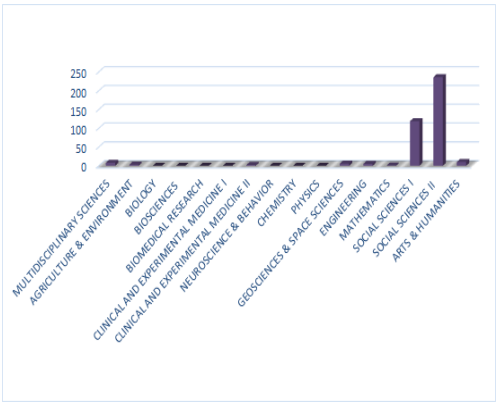


Figure 4. Number of publications categorized by Glänzel and Schubert classification.

[1]UK left EU on February 1, 2020.

Conclusion

The bibliometric analysis of the EaP's 15-year journey reveals a dynamic and evolving academic interest in this initiative. The involvement of various countries, including significant contributions from EU member states and other non-EU countries, demonstrates the global relevance of the EaP. This growing scholarly attention serves as an important reminder of the EaP's significance for both regional stability and deeper integration into European structures.

References

Laure Delcour, Shaping the Post-Soviet Space? EU Policies and Approaches to Region-Building, 2011.

George Petrakos, Maria Tsiapa and Dimitris Kallioras, Regional inequalities in the European Neighborhood Policy countries: The effects of growth and integration, Environment and Planning C: Government and Policy 2015, volume XX, pages 1 –19

Elena A. Korosteleva, Change or Continuity: Is the Eastern Partnership an Adequate Tool for the European Neighbourhood? International Relations 25(2), 2011, 243–262.

Eastern Partnership, European Union External action, https://www.eeas.europa.eu/eeas/eastern-partnership_en

Glänzel W., Schubert A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics 56, 357–367.

https://doi.org/10.51408/issi2025_201

A Framework for Analyzing Identification Funds in the Social Sciences under the Perspective of Country Mentions: An Example of China and the United States

Changcheng Xue[1], Kaiwen Shi[2], Hongyu Wang[3], Xiaoguang Wang[4]

[1]*changcheng_xue@163.com*, [3]*hongyuwang@whut.edu.cn*
Wuhan University of Technology, Wenzhi Street, Hongshan District, Wuhan (China)

[2]*shikaiwen@whu.edu.cn*, [4]*wxguang@whu.edu.cn*
School of Information Management, Wuhan University, Bayi Street, Wuchang District, Wuhan (China)

Introduction

Funded papers are those produced with research funds from government departments, funding organizations, and enterprises. Identification fund is a fund that specializes in funding as well as large-scale funding for research in the social science field. Research productivity grants support various scientific fields (Marcelo Perlin, 2024). Evaluating research funding effectiveness is valuable for policymakers (Guiyan Ou, 2024). They're interested in the effectiveness of competitive grant models (Alberto Corsini, 2023). Assessing academic research funding is tough due to diverse sources. So, identifying key aspects is crucial (Mike Thelwall, 2023). EU FPS funding is skewed (Fredrik Niclas Piro, 2024). The current research is mainly from the perspective of research managers, involving the management decision-making and performance evaluation of research funds, and lacks the excavation of the research content of fund support from the national level, especially for the field of humanities and social sciences.

This study presents the concept of Identification Fund in the field of social sciences, and its analytical framework. The framework analyzes the topics of papers funded by the Social Science Field Identification Fund at the content level of scientific knowledge carriers (research papers), and is able to observe the main research content of different research subjects at the national level when they are mentioned to each other. Taking China and the United States as the mentioned subjects, it reveals how the American, British, German, and European focus on China and the United States is similar and different and how the research topics evolve.

Methods

First, screen paper data from the WOS database and determine the list of identified funds with the fund's official website. Then, select funded papers by identified funds. Next, use NEViewer (Wang X, et al. 2014), VOSviewer software and big data methods to analyze literature data.

We chose papers from the Web of Science database, screening 2,437,656 papers in 49 social - science fields from 2014 - 2023, and identified 22 funds. Using the database's advanced search, we input FO, WC, and PY to search for each fund, extract paper fields and topics, and keep only one for repeated fields. As shown in Fig.1.

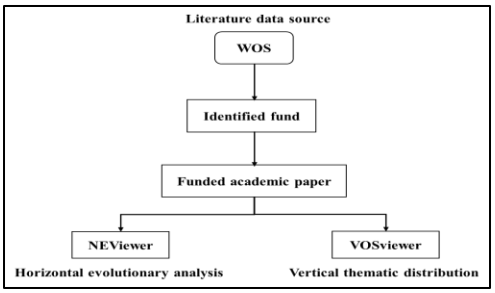


Figure 1. Technical Roadmap.

Result

Use VOSviewer to draw a distribution map with keywords of funded papers from different countries and extract main topics by key

keywords, as in Fig.2 and Fig.3. Use NEViewer for horizontal evolution analysis of funded papers, as shown in Fig.4 and Fig.5.

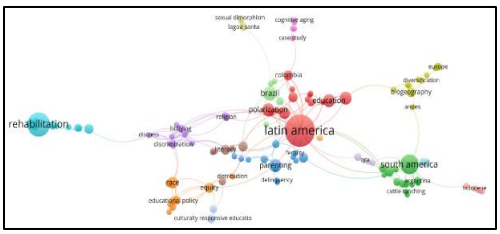


Figure 2. Distribution Map of Germany's Focus on US topics.

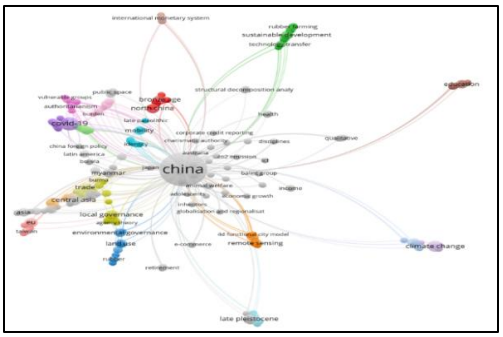


Figure 3. Distribution Map of Germany's Focus on Topics Related to China.

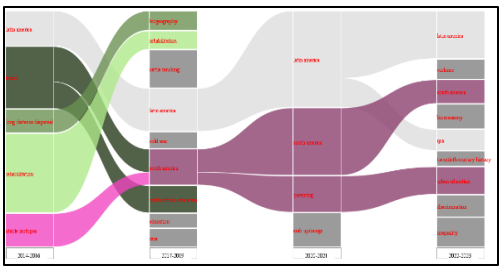


Figure 4. Evolution Analysis Chart of Germany's Focus on US topics.

From 2014 - 2023, research topics evolved dynamically in regions, disciplines, and social focus. Geographically, it shifted from local areas to Latin America, refined to South America, and deepened around them. In disciplines, it changed from practical fields to multi - disciplines, then to macro - social and urban fields, and finally to interdisciplinary areas. Socially, the focus moved from local social structure to industries and cultural communication, then to social development and urban construction, and finally to macro - social issues.

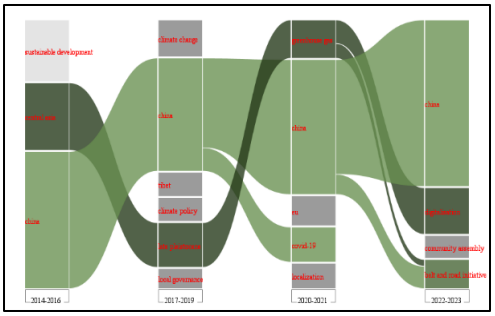


Figure 5. Evolution Analysis Chart of Germany's Focus on Chinese Topics.

From 2014 to 2023, relevant research topics evolved dynamically in development trends, research focuses, and policy correlations. For development trends, it went from emphasizing sustainable development in 2014 - 2016, to focusing on climate change in 2017 - 2019, then to energy issues in 2020 - 2021, and centered on water resources and deepened in 2022 - 2023.In research focuses, it was on resource management in 2014 - 2016, covered land policies etc. in 2017 - 2019, concentrated on ecological balance in 2020 - 2021, and involved interdisciplinary aspects like environmental security in 2022 - 2023.Regarding policy correlations, it related to local resource policies in 2014 - 2016, echoed national climate policies in 2017 - 2019, was associated with regional energy planning in 2020 - 2021, and linked closely to global water resources management policies in 2022 - 2023.

Conclusion

This study proposes a framework for analyzing the identification fund in the field of social sciences from the perspective of country mentioning. By analyzing the papers funded by the Identification Fund in the field of social sciences horizontally and vertically, we can obtain the main research topics and their evolution process of different scientific research subjects when mentioning other countries, which can help to grasp the scientific research trends in the field of social sciences at a higher level.

Acknowledgments

This work was funded by the National Natural Science Fund of China (No. 71874129).

References

Marcelo Perlin. (2024). The determinants and impact of research grants: The case of Brazilian productivity scholarships. Journal of Informetrics, Volume 18.

Guiyan Ou. (2024). Effects of research funding on the academic impact and societal visibility of scientiffc research. Journal of Informetrics, Volume 18.

Alberto Corsini. (2023). Does grant funding foster research impact? Evidence from France.Journal of Informetrics, Volume 17.

Mike Thelwall. (2023). What is research funding, how does it infuence research, and how is it recorded? Key dimensions of variation. Scientometrics, 128:6085–6106.

Fredrik Niclas Piro. (2024). Regional and sectoral variations in the ability to attract funding from the European Union's Seventh Framework Program and Horizon 2020. Scientometrics, 129: 1493–1521.

Wang X, Cheng Q, Lu W. Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. Scientometrics, 2014, 101(2):1253-1271.

https://doi.org/10.51408/issi2025_172

A look behind metrics for knowledge integration: Some notable cases

Pei-Shan Chi[1], Wolfgang Glänzel[2]

[1]*peishan.chi@kuleuven.be*, [2]*wolfgang.glanzel@kuleuven.be*
Faculty of Economics and Business, ECOOM, KU Leuven, Naamsestraat 61, 3000, Leuven (Belgium)

Introduction

In earlier papers, we have used the analysis of cited references to study cognitive aspects of interdisciplinarity (IDR) in scientific research. We assumed IDR being an expression of knowledge integration that can be traced by analysing cited references, which in turn are considered a form of use of scientific information in the framework of documented scholarly communication. Yet, measure based on cited references tend to overestimate cognitive links in favour of methods and instruments used (e.g., Glänzel & Thijs, 2017). The same applies to IDR measures, if those are based on citation links. In particular, we found that while, at the nano level, the distinction between IDR and multi-disciplinarity is straightforward, the distinction between IDR and cross-disciplinarity (CDR) remains a challenge. We used variety and disparity measures to describe important characteristics of IDR, but found striking examples, notably of high disparity, in which the extent of knowledge integration is questionable. Papers in archaeology and religion, in which advance imaging technologies or instruments were used and referred to in the bibliography without true integration of the underlying knowledge into the research, may just serve as an example. The citers of these studies typically remained in the field of archaeology or religion. This forced us to assume that the role of users (citers) also play an important role in the understanding of knowledge integration.

Basic approach

We proceed from the assumption that true knowledge integration takes place if and only if some new research is established combining sources from different disciplines or is used in research of one of the disciplines referred to, but information is used by other fields and not only by this discipline. Figure 1 gives an illustration of three typical examples. Top-left corner: two disciplines are cited, but only one of those cite the results. Top-right corner: the same references, but both disciplines are using the results. Bottom: knowledge from two disciplines is combined in a different discipline and cited there. The first case may not reflect true integration of knowledge.
This leads us directly to the following important question.

- What typical aspects need to be considered to improve the meaningfulness of IDR metrics for cited and citing literature?

In the following, we propose some metrics-based method to answer this question and give examples of specific cases of knowledge use vs. integration.

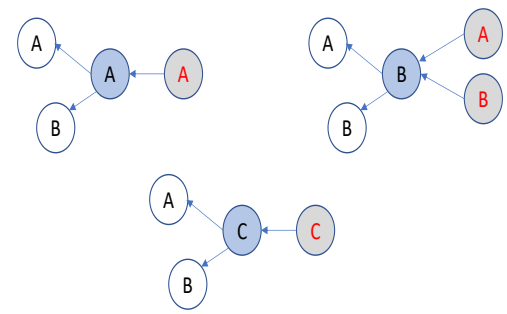


Figure 1. Three examples of information use from and in different disciplines (blue: source, white: references, grey: citations).

Methodology

As already proposed, e.g., by Stirling (2007) and Leinster & Cobbold (2012), we will apply two standard measures commonly used in bibliometric research on IDR: variety (V_S) and disparity (D_S). In particularly, we will use V_S and D_S for all source items in conjunction with *Characteristic Scores and Scales* (CSS) classes to obtain scale-independent measures (cf. Glänzel & Debackere, 2022). To identify potentially problematic cases, we will select papers with outstanding disparity but low variety because these use knowledge from few but very distant disciplines. In a second step, we retrieve all citations to these source items. Both papers and citations are taken from recent volumes of Clarivate Web of Science Core Collection. For each source item, we also calculate the V_C and D_C values based on citing items. Finally, we determine the profile similarity of reference items of each source paper and the citing set of papers (s_{rc}). While V_C and D_C are used to further separate cases, the last indicator will be used to help answer the research question.

Results

To trace the appearance of the cases such as in the model sketched in Fig. 1 based on a systematic approach, we have set a filter on the D_S of all papers indexed for the year 2019 in the Web of Science Core Collection (WoS). We focused on the highest CSS class of disparity (cf. Glänzel & Debackere, 2022), i.e., we selected those papers that used information from different but not related fields. In turn, we focussed on two cases of patterns of citations received by these papers, high and low V_C and reasonable D_C values, if V_C is large. Finally, we use the (dis-)similarity of citation and reference sets of each paper as kind of validation of our results but also to detect outliers. A low similarity between reference and citation profiles would be unrealistic but we can expect moderate to large similarity according as use of knowledge in citations differs from that used in references. Thus, instead of just presenting statistics, we intended to lock "behind" these cases to better understand the mechanism of knowledge use, diffusion and integration.

In the following, we give some examples together with their indicator values. Before we give a small example set, we point to two interesting archetypes of IDR-related knowledge diffusion. DOI: 10.1371/journal.pone.0239831 ("*The length of a scroll: Quantitative evaluation of material reconstructions*") in ancient religion cited literature from religion and physical chemistry (imaging technology) with high D_S, but results are apparently only relevant for religion (low V_C). The single-authored DOI: 10.1093/isd/ixz006 in entomology ("*A systematist's guide to estimating Bayesian phylogenies from morphological data*") tells against the myth that IDR requires co-operation of researchers with different professional background. Both cited and citing papers represent sets of broad and similar subject profiles including entomology, evolutionary biology, genetics heredity, ecology, palaeontology, zoology, anatomy, morphology, and mathematics-/computer science in the references.

Table 1. A sample representing ten IDR/CDR papers with different disparity/variety values of their references and citations with moderate to strong similarity of cited and citing literature.

DOI	D_s	D_c	V_s	V_c	s_{rc}
10.1016/j.nimb.2018.05.002	3.77	3.89	6.48	7.36	0.89
10.1089/ast.2017.1746	3.08	4.27	5.18	6.32	0.90
10.1016/j.envres.2018.09.039	3.51	4.37	6.53	9.45	0.82
10.1080/09296174.2017.1405719	3.62	2.52	6.89	4.00	0.59
10.1016/j.enpol.2018.07.040	3.18	2.42	4.57	4.17	0.48
10.1039/c8an01059e	4.49	2.48	12.25	5.56	0.79
10.1177/0022429418799362	3.11	1.06	6.44	3.00	0.70
10.1039/c8an01526k	3.25	3.41	8.05	10.75	0.91
10.1016/j.jvs.2018.04.029	2.96	3.61	5.72	10.89	0.75
10.1016/j.saa.2018.09.051	3.77	3.26	8.41	9.92	0.45

Table 1 gives a small part of records with interesting quadruples of indicator values in which we found remarkable cases. Some will be discussed here. DOI: 10.1080/09296174.2017.1405719 ("*The Stylometric Impacts of Ageing and Life Events on Identity*") in quantitative linguistics with moderate profile similarity of cited and citing literature and strong disparity uses literature from a large range of subjects in neuroscience and behavioral sciences, psychology, linguistics, literature, ecology, computer science and some related fields, with citation impact on telecommunications, computer science, linguistics, electrical and electronic engineering. Cited and citing literature show different foci. DOI:

10.1016/j.enpol.2018.07.040 ("*Costs and benefits of saving unprofitable generators: A simulation case study for US coal and nuclear power plants*") uses knowledge from environmental science, electrical and electronic engineering, economics, cardiology, energy and fuels, computer science, while it impacts on economics, energy and fuels, chemical engineering, environmental sciences and mathematics. Again, the different foci lower similarity of profiles. DOI: 10.1177/0022429418799362 ("*Music Performance Anxiety and Perceived Benefits of Musical Participation Among Older Adults in Community Bands*") combines knowledge from music, psychology, neurosciences, education, gerontology, medical sciences, health care and impacts the same disciplines however with a somewhat narrower scope.

Conclusions

We have briefly discussed five noticeable cases obtained from the application of interdisciplinarity metrics. These examples show that it is worthwhile looking "behind" the indicators to correctly interpret IDR-related phenomena. In the analysis of larger sets, we found documents that used relevant literature without true integration of knowledge. Others produced knowledge outside used literature. In future research, we will develop further filters to detect papers and distinguish types with typical and atypical patterns of knowledge integration and diffusion on a largescale. We expected to give further insight into the mechanisms of creating new knowledge relevant even beyond the disciplinary scope of literature used for the research.

References

Glänzel, W. & Debackere, K. (2022), Various aspects of interdisciplinarity in research and how to quantify and measure those. *Scientometrics*. 127(9), 5551–5569.

Glänzel, W. & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset. *Scientometrics*. 111(2), 1071–1087.

Leinster, T., & Cobbold, C.A. (2012). Measuring diversity: The importance of species similarity. *Ecology*, 93(3), 477–489.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719.

https://doi.org/10.51408/issi2025_189

AI-Powered Evaluation of Peer Review Quality: A Case Study of eLIBRARY.RU

Dmitry Kochetkov[1], Denis Kosyakov[2], Irina Lakizo[3], Viktor Glukhov[4], Andrey Guskov[5]

[1]*kochetkov@elibrary.ru, d.kochetkov@cwts.leidenuniv.nl*
Scientific Electronic Library LLC, Nauchny Proezd 14A block 3, 117246 Moscow (Russian Federation)
Centre for Science and Technology Studies, Leiden University, Kolffpad 1, 2333 BN Leiden (The Netherlands)

[2]*kosyakov@sscc.ru, d.kosyakov@riep.ru,* [5]*guskov@sscc.ru, a.guskov@riep.ru*
Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Ac. Lavrentieva ave. 6, Novosibirsk (Russian Federation)
Russian Research Institute of Economics, Policy and Law in Science and Technology, Dobrolubova Str. 20A, Moscow (Russian Federation)

[3]*i.lakizo@riep.ru*
Russian Research Institute of Economics, Policy and Law in Science and Technology, Dobrolubova Str. 20A, Moscow (Russian Federation)

[4]*olunid@elibrary.ru*
Scientific Electronic Library LLC, Nauchny Proezd 14A block 3, 117246 Moscow (Russian Federation)

Introduction

eLIBRARY.RU is the largest Russian electronic library of scientific publications and home to the Russian Index of Science Citation (RISC) and highly selective Russian Science Citation Index (RSCI). One of the challenges we face in the expert evaluation of review quality and journal policies is the shortage of qualified experts. A potential solution to this problem is the use of *generative artificial intelligence* (GenAI) technologies to assess the quality of reviews. Recent studies cautiously evaluate the potential of GenAI in scientific peer review. For example, AI tools can assist in the initial screening of articles, plagiarism detection, and reviewer matching, potentially saving millions of working hours (Checco et al., 2021). However, concerns remain about biases and ethical implications (Shcherbiak et al., 2024). Seghier (2025) advocates for the gradual integration of AI into the peer review process under human oversight, emphasizing the importance of transparency, accountability, and robust safeguards. At the same time, *the potential of AI technologies for evaluating review quality remains largely unexplored*.

The goal of this study is to address the question of *whether AI-based evaluation of journal review quality is feasible at the current level of technological development.* This report presents preliminary findings based on a test sample of 240 reviews.

Data and Methods

To assess peer review quality, we created a test sample by selecting four diverse disciplines (*Economics & Business, Information & Computer Science, Physics & Mathematics, and Medicine*) to test AI versatility across different research types. Within each discipline, we chose two journals representing high-impact (top 1-500) and mid-tier (1501-2000) rankings in the *Science Index*[1], randomly selecting 30 review reports

[1] Science Index is a composite journal ranking on eLIBRARY.RU.

from each journal. This approach ensured a diverse sample spanning methodological approaches and journal prestige levels.

The selected reviews were evaluated using two sets of criteria. The first set, based on Russian Science Citation Index parameters, assessed *depth, usefulness, rigor, and clarity*. The second set adapted the *Review Quality Instrument (RQI)* (van Rooyen et al., 1999), evaluating eight aspects: *research importance, originality, methods, presentation, comment constructiveness/substantiation, result interpretation, and overall quality*. Each criterion was scored on a detailed 5-point Likert scale. GPT-4 was employed via API to assign scores and provide justifications, specifically referencing the review text for the RQI criteria. The process ensured no disclosure of personally identifiable information.

Results

The results of the analysis based on *Criterion Set 1* are presented in Figure 1. Journals are categorized by subject area: Economics and business (eco), Information and computer science (info), Physics and mathematics (phys), and Medicine (med), as well as by their ranking range in the Science Index (SI) – 1-500 (index 1) or 1501-2000 (index 2).

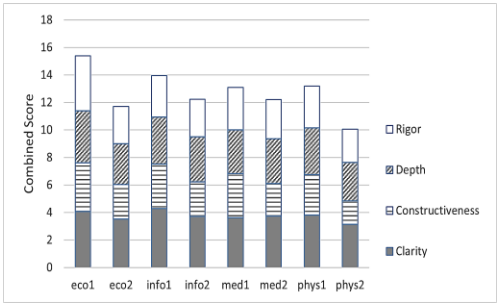


Figure 1. Average scores by journals' categories according to Criterion Set 1.

The quality of reviews in journals across all disciplines was higher for those in the SI 1-500 range compared to those in the 1501-2000 range. This finding indirectly supports the hypothesis of a correlation between bibliometric indicators and the quality of editorial policies, particularly peer review.

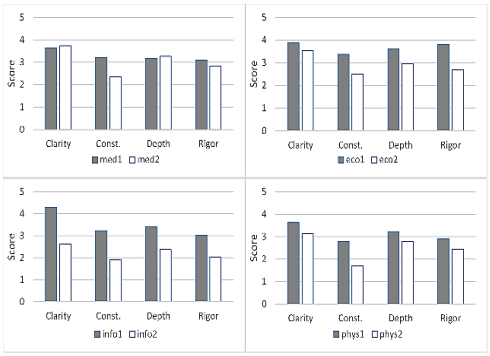


Figure 2. Comparative analysis of the average scores by criteria and journals' category according to Criterion Set 1.

For *Clarity* and *Depth* criteria, we see a superiority of mid-tier journal scores over high-impact journal scores in *Medicine* (*Figure 2)*. In other disciplines, review scores for all four criteria are weaker for mid-tier journals.

The application of *Criterion Set 2* yielded slightly different results (*Figure 3*). In this case, the difference between journals in the two ranking tiers was less pronounced in economics and business. Moreover, the medical journal in the 1501-2000 range performed slightly better than its counterpart in the 1-500 range. In contrast, the advantage of high-impact journals is more pronounced in the other two areas.

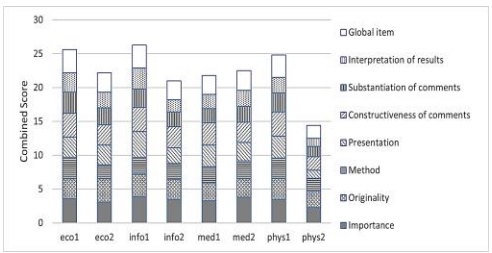


Figure 3. Journals' scores according to Criterion Set 2.

To analyze in detail the results that do not fit the intended picture, we compared the scores of the journals for each criterion (*Figure 4*). The mid-tier medical journal outperformed the high-impact journal in all but two criteria: *Presentation*, and *Constructiveness of comments*. The superiority of mid-tier journal is also observed in the field of *Economics and business* in terms of *Originality* and to a lesser extent in terms of *Presentation*. The most

significant difference was observed for criterion *Importance*.

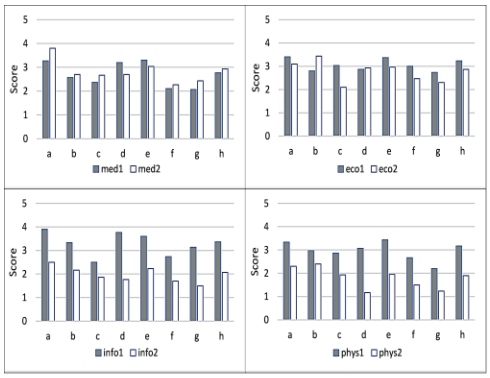


Figure 4. Comparative analysis of the average scores by criteria and journals' category according to Criterion Set 2. Criteria: a – importance, b – originality, c – method, d – presentation, e - constructiveness of comments, f - substantiation of comments, g - interpretation of results, h- global item.

Competing Interests

Dmitry Kochetkov and Viktor Glukhov are Deputy Directors of Scientific Electronic Library LLC, the operator of eLIBRARY.RU, RISC, and RSCI.

References

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1), 25. https://doi.org/10.1057/s41599-020-00703-8

Seghier, M. L. (2025). AI-powered peer review needs human supervision. *Journal of Information, Communication and Ethics in Society*, 23(1), 104–116. https://doi.org/10.1108/JICES-09-2024-0132

Shcherbiak, A., Habibnia, H., Böhm, R., & Fiedler, S. (2024). Evaluating science: A comparison of human and AI reviewers. *Judgment and Decision Making*, 19, e21. https://doi.org/10.1017/jdm.2024.24

van Rooyen, S., Black, N., & Godlee, F. (1999). Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts. *Journal of Clinical Epidemiology*, 52(7), 625–629. https://doi.org/10.1016/S0895-4356(99)00047-5

https://doi.org/10.51408/issi2025_203

Analysis of compliance with the FAIR principles in Education Science

Andrea Sixto-Costoya[1M], Adolfo Alonso-Arroyo[2], Luiza Petrosyan[3], Rafael Aleixandre-Benavent[4], Rut Lucas-Domínguez[5]

[1]*andrea.sixto@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia
Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of Social Work and Social Services, University of Valencia (Spain)

[2]*adolfo.alonso@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia
Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)

[3]*luipet@inf.upv.es*
Polytechnic University of Valencia. University Institute of Pure and Applied Mathematics (IUMPA) (Spain)

[4]*rafael.aleixandre@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia
Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Ingenio (CSIC-Polytechnic University of Valencia) (Spain)

[5]*rut.lucas@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia
Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)
CIBERONC (Spain)

Introduction

The roles of accountability in sharing Research data and the ability to reproduce experiments have already been widely pointed out. To operationalize the practice of data sharing, the so-called FAIR principles, which stand for "findable," "accessible," "interoperable" and "reusable," were published in 2016 (Wilkinson et al., 2019). Without compliance with these principles, data quality can be so low that it becomes useless due to the difficulty of understanding it. As was pointed out, data quality is commonly conceived as a construct that is defined by the extent of its usefulness (Brennan, 2017).

In the field of education, raw research data are crucial because they allow for a better understanding of research on educational interventions and learning, which is considered one of the fundamental pillars of human, social and economic development. Their quality must be guaranteed to avoid the risk of misinterpretation or bias. Therefore,

our objective is to assess the quality of a set of educational data sets.

Methods

The methodology used in this study consists of three stages:

1. Capturing datasets on Education Sciences. A search equation was designed to retrieve datasets related to Education. The search was conducted in OpenAlex database where the term "Education" appeared in the Subfield OR Keywords fields.
2. Downloading the records. The total recovered records (datasets) were N=65,199. Looking at the repositories in which they are included, there are 223 different repository variables. For this study, the generalists Zenodo and Figshare repositories were selected. These records were downloaded in.txt format and processed with our *Bibliometricos* software. Once the information was parsed and organized, it was necessary to know the unique identifier (DOI) of the total records. This identifier is needed in the next step of the methodology.
3. FAIR evaluation of datasets with the F-UJI tool. The FAIR assessment for these datasets was performed with the F-UJI tool (https://www.f-uji.net/), that evaluates research data objects, which is a REST API using OpenAPI Specification from a remote server, published under an open-source MIT licence. It is based on aggregated metadata, including metadata embedded in the landing page and metadata retrieved from a DOI. The outcomes of such evaluations yield diverse scores pertaining to the metadata of data and datasets, with 16 metrics distributed across four principles: findability (5 metrics), accessibility (3 metrics), interoperability (3 metrics), and reusability (5 metrics) (Devaraju et al., 2022). This methodology was used and validated previously in Petrosyan et al. (2023) and Sixto-Costoya et al. (2025).

Results

After the analysis through the F-UJI tool, we obtained information about the level of FAIRification of the 4,642 DOI belonged to datasets in the Education Sciences area. Of them, 1,772 belonged to Zenodo and 2,483 to Figshare.

Through the report obtained by the F-UJI tool, we can observe the degree of compliance with the FAIR principles of the Education datasets in the two repositories studied. The first thing we can observe is that it is in Figshare where the best FAIR percentage was obtained in terms of average (Table 1). Furthermore, it is also observed that not only does Figshare achieve better mean FAIR compliance than Zenodo, but when looking at the distribution of scores, none of the Zenodo datasets achieve even 50% compliance (Figures 2 and 3).

Table 1. Average of the percentage of compliance with the FAIR principles obtained by the two repositories analysed.

REPOSITORIES	ANALISED DOIS (num)	TOTAL FAIR %
FIGSHARE	2,483	56.86
ZENODO	1,772	43.30

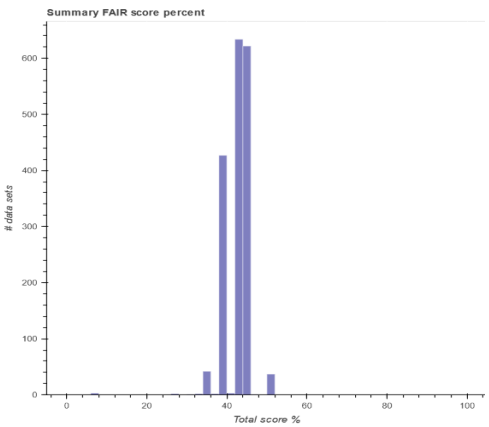


Figure 1. Percentage of compliance with FAIR principles in Zenodo repository.

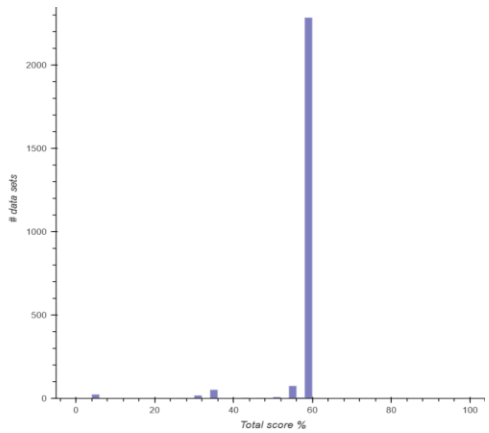


Figure 2. Percentage of compliance with FAIR principles in Figshare repository.

A similar result is observed when the percentage of compliance with the FAIR principles is looked at separately (Table 2). Overall, Figshare is better on three principles, only narrowly beaten by Zenodo on the Accessible principle. It is noteworthy that this Accessible principle is the lowest scoring of the two repositories, and it should be noted that only Findable achieves more than 50% compliance.

Table 2. Average of the percentage of compliance with each of the FAIR principles in a differentiated manner obtained by the two repositories analysed.

REPOSITORIES	F	A	I	R
FIGSHARE	84.3	33.1	47.8	48.42
ZENODO	65.4	33.3	31.3	35.6

Conclusions

The preliminary results of our study showed the issues that remain to be resolved, especially in relation to the FAIR principle of Accessibility, but also to Interoperability and Reusability. However, it is important to note that Findable is a principle that, at least in the two repositories studied, is acceptable.

Further in-depth analysis of the causes and possible solutions for the improvable score of the other three principles is crucial for the development of data sharing practices in Educational Sciences. This is an area that has a direct impact on the well-being of citizens and whose improvement in terms of research is necessary to make faster progress.

Acknowledgments

This work has benefited from an aid from the Ministry of Science and Innovation of the Government of Spain. State Research Agency. Ecological and Digital Transition Projects 2021. TED2021-131057B-I00.

References

Brennan, R. (2017). Challenges for Value-driven Semantic Data Quality Management. *Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 1*. https://doi.org/10.5220/0006387803850392

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J. & White, A. (2022). FAIRsFAIR Data Object Assessment Metrics (0.5). In *Zenodo* (Issue October). https://doi.org/10.5281/zenodo.6461229

Petrosyan, L., Aleixandre-Benavent, R., Peset, F., Valderrama-Zurián, J. C., Ferrer-Sapena, A. & Sixto-Costoya, A. (2023). FAIR degree assessment in agriculture datasets using the F-UJI tool. *Ecological Informatics*, *76*(May). https://doi.org/10.1016/j.ecoinf.2023.102126

Sixto-Costoya, A., Ferrer-Sapena, A., Aleixandre-Benavent, R., Peset, F., Valderrama Zurián, J. C. & Petrosyan, L. (2025). The compliance to FAIR principles of shared data in addiction. *Scientometrics*, *0123456789*. https://doi.org/10.1007/s11192-024-05227-5

Wilkinson, M. D., Dumontier, M., Sansone, S. A., Bonino da Silva Santos, L. O., Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M. E. & Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data*, *6*(1), 1–12. https://doi.org/10.1038/s41597-019-0184-5.

https://doi.org/10.51408/issi2025_208

Attempts to Enable Generative AI for Topic Recognition: A Case Study of ChatGPT

Wenting Tang[1], Wen Lou[2]

[1]twt2543535795@163.com,[2]wlou@infor.ecnu.edu.cn
East China Normal University, 3663 Zhongshan North Road, 200062, Shanghai (China)

Introduction

In recent years, the application of generative AI has seen growing use in NLP tasks like keyword extraction, entity recognition, and translation (Lu et al., 2024), yet its role in topic recognition remains underexplored.

Traditional topic models like LDA and PLSA build thematic spaces via word co-occurrence matrices, often causing semantic ambiguity and feature sparsity in theme inference. In contrast, generative AI develops deep contextual semantic representations through massive corpus pre-training, enabling accurate identification of implicit themes and effective mitigation of theme recognition bias.

This study aims to explore the application of generative AI, specifically ChatGPT, in topic recognition in scientific literature. The study attempts to achieve efficient topic recognition through two different strategies and compares two strategies with machine learning methods. The study evaluates the advantages and limitations of generative AI in topic recognition, providing richer offering empirical insights for its practical use in literature analysis.

Methodology

Strategy One: Topic Recognition Based on Excel Files

This strategy enables ChatGPT to process large metadata from Excel files. Using PubMed as the source, the study filters medical literature published between 2000 and 2020, with article types including Clinical Trial, Meta-Analysis, and Randomized Controlled Trial. Using web scraping, key data like titles, abstracts, keywords, and publication dates are extracted and formatted into an Excel file with 17,000 records.

For topic recognition, this strategy attempts to use ChatGPT to perform topic recognition based on the BERT model (Sawant et al., 2022). The strategy provides ChatGPT with a basic explanation of the BERT framework and uses specific instructions to guide ChatGPT in performing BERT-based clustering. Specifically, ChatGPT is instructed not to directly provide the BERT model code but to encode each piece of metadata using the BERT model, extract its semantic features, and apply a clustering algorithm to group similar literature into categories for topic aggregation and recognition.

Strategy Two: Topic Recognition Based on Abstract Content

This strategy involves directly inputting the literature titles and abstracts into ChatGPT in the form of a dialogue to perform topic recognition. Specifically, this strategy guides ChatGPT to follow the steps of the DBSCAN model for topic clustering (Luchi & Rodrigues, 2019).

The strategy first instructs ChatGPT to removestopwords and numbers, normalize word forms, and construct a document vocabulary list from the abstracts. It then calculates TF-IDF and cosine similarity to assess topic similarity. With defined ε and MinPts, it classifies metadata into core, border, and noise points, further organizing the data into topic categories to provide insights into potential research topics.

Results and Discussion

Discussion of Strategy One: Topic Recognition Based on Excel Files

Table 1. BERT Keywords VS ChatGPT Keywords.

	Topic Feature Keywords Identified by ChatGPT	Topic Feature Keywords Identified by BERT
Topic 1	hypertension, treatment, risk	hypertension, amlodipine, antihypertensive
Topic 2	cancer, lung, factors	acupuncture, rehabilitation, stroke
Topic 3	infection, H. pylori, gastric	nutrition, parenteral, enteral
Topic 4	community, effectiveness, intervention	propofol, anesthesia, dexmedetomidine
Topic 5	clinical, randomized, controlled	rectal, laparoscopic, anastomosis

Both ChatGPT and the BERT model identified topics related to hypertension treatment and antihypertensive drugs. However, ChatGPT emphasized a broader evaluation of "effects" and "risks," while BERT concentrated on specific medications like "amlodipine" and their impact on blood pressure control. BERT's topics were more detailed, exploring specific treatments, whereas ChatGPT identified overarching themes about hypertension treatment effectiveness.

For other topics, there was minimal overlap between ChatGPT and BERT. ChatGPT's themes were broader, suitable for detecting trends in large datasets, while BERT excelled in semantic accuracy and context, particularly in recognizing technical terms and treatment methods.

According to Bougioukas 's literature review (Bougioukas et al., 2021), keywords like "systematic review" and "study" appear most often, aligning with ChatGPT's Topic 1. Medical terms such as "acupuncture," "cancer," and "effectiveness" match BERT's Topic 2 and ChatGPT's Topics 1, 2, and 4. This suggests bibliometric methods produce research topics semantically and topically similar to those from generative AI.

Discussion of Strategy Two: Topic Recognition Based on Abstract Content

In topic recognition based on abstract content, although effective topic clustering was achieved by following the steps of the DBSCAN model, practical challenges still arose.

First, the issue of selecting ε and MinPts. The key to DBSCAN lies in selecting ε and MinPts. Manual tuning often requires multiple trials to optimize clustering, during which ChatGPT may produce memory errors—like fabricating cosine similarities between fictional documents—causing result deviations and reducing topic recognition accuracy.

Second, the issue of accurately comparing numerical values. Since DBSCAN's reliance on cosine similarity involves comparing small decimals. ChatGPT may misjudge values with varying decimal places (e.g., seeing 0.3 as smaller than 0.11), leading to misclassification of core points and distorted clustering.

Third, there is the issue of input and output word count limits. While batch processing helps mitigate word count restrictions, merging data from different batches may exceed the system's capacity, reducing efficiency and impacting the stability of the results.

Conclusion

This study explores the application of generative artificial intelligence in topic recognition of medical literature through two strategies: Excel files and abstract content. In the Excel-based approach, only one ChatGPT topic aligned with BERT's; BERT captured finer details, while ChatGPT identified broader themes but missed semantic nuance. The abstract-based strategy enabled effective clustering but faced issues with parameter tuning, numerical precision, and word count limits.

Overall, generative AI holds promise for topic recognition but requires further optimization for large-scale data and semantic precision. Future work will integrate traditional methods

with generative AI to enhance efficiency and accuracy.

Acknowledgments

This work was supported by the Independent Research Project of Key Laboratory of Frontier Theory and Application of Statistics and Data Science of Ministry of Education, Identification and Application of Key Technologies in Medical Research Empowered by Digital Intelligence [KLATASDS2406], and by the Shanghai Planning Office of Philosophy and Social Science Project, Research on the Theoretical Framework and Implementation Path of Intelligent Sharing of Health Science Data[2024BJC005].

References

Bougioukas, K. I., Vounzoulaki, E., Mantsiou, C. D., et al. (2021). Global mapping of overviews of systematic reviews in healthcare published between 2000 and 2020: a bibliometric analysis. Journal of Clinical Epidemiology, 137, 58–72.

Luchi, D., & Rodrigues, A. L. (2019). Sampling approaches for applying DBSCAN to large datasets. Pattern Recognition Letters, 117, 90-96.

Lu, W., Liu, Y. P., Shi, X., et al. (2024). Academic text mining driven by large models: Construction of inference-end instruction strategies and capability evaluation. Journal of the China Society for Scientific and Technical Information, 43(08), 946-959.

Sawant, S., Yu, J., Pandya, K., et al. (2022). An enhanced BERTopic framework and algorithm for improving topic coherence and diversity. IEEE 24th International Conference on High Performance Computing & Communications.

https://doi.org/10.51408/issi2025_173

Automating Reproducible Bibliometrics with the Open Research Converter

Jack H. Culbert[1], Philipp Mayr[2]

[1]*jack.culbert@gesis.org*, [2]*philipp.mayr@gesis.org*
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667, Cologne (Germany)

Abstract

The Open Research Converter[1] (ORC) is an open-source tool that allows users to convert Digital Object Identifiers into OpenAlex[2] Work IDs and/or retrieve full bibliometric records from OpenAlex. In this poster paper, we introduce the ORC and show its main application: the generation of open and sharable bibliometric datasets, future development plans and a short analysis of usage patterns so far.

Introduction

Bibliometric and Scientometric studies which involve bibliometric data taken from proprietary databases (such as the Web of Science (WoS) or Scopus) suffer from a lack of openness, transparency, and reproducibility as researchers are not permitted to freely share and publish the underlying data from their analyses. Workarounds such as "we searched the [*query terms q*] and exported *n* records from WoS version *m*" have been utilised by the community but remain difficult a barrier to reproducibility as the underlying dataset from the study is unavailable.

Reproducibility in Bibliometrics and Scientometrics has been previously studied resulting in (Velden et al., 2018), and the current data sharing and publishing restrictions with the commercial providers are not likely to change in the short term. Consequently, bibliometric research based on WoS and Scopus data is likely to remain unreproducible and lacks the transparency which is required for Open Science research.

OpenAlex (Priem et al., 2022) was released in 2022 and is an open-source bibliometric database which releases its data under a maximally permissive license (CC0 1.0 Universal), which enables researchers to share their datasets. However, frictions for bibliometricians exist, including adapting to the website interface, the technical knowledge to utilise the API or raw data (provided by OurResearch[3] as a monthly approximately 300GB JSON snapshot), and a healthy suspicion of the quality of the bibliometric dataset.

The Open Research Converter (Culbert et al., 2024) was primarily designed to assist bibliometricians with the lattermost friction, allowing them convert DOIs from within their dataset to OpenAlex Work IDs, which can then be shared alongside their publications – increasing reproducibility and openness within the Scientometrics community and elsewhere. Since then, following community feedback at the Nordic Workshop on Bibliometrics & Research Policy 2024 (Culbert, 2024), we have been developing new features as detailed below.

[1] orc-demo.gesis.org
[2] openalex.org
[3] ourresearch.org



Figure SEQ Figure * ARABIC 1 – The Open Research Converter Interface, overlaid with a snippet from the full record output.

Open Research Converter

The ORC is a containerised Python application with a JavaScript frontend which allows researchers to input Digital Object Identifiers (DOIs) manually or upload a csv and returns either the OpenAlex Work ID or the full bibliographic record in csv format. (The codebase is available via Github.) [4]

The ORC provides bibliometrics researchers with the ability to use DOIs to identify the records in OpenAlex which match those in other databases. This approach has its limitations, as explored in (Vieira & Leta, 2024), such as missing or duplicated DOIs, and therefore we are working on a fuller approach which incorporates other publication metadata into the matching process.

DOIs accompany most bibliometric records in both proprietary academic databases such as the Web of Science (WoS), Scopus, and Dimensions and open databases such as PubMed, ArXiv, Semantic Scholar, OpenAIRE and OpenAlex. The degree of overlap and number of records without a DOI in WoS, Scopus and OpenAlex (and thereby the accuracy of this method) was explored in (Culbert et al., 2024).

The ORC backend is capable of processing over 300,000 records in a single request and is only limited by the size of the input CSV allowable in the frontend, to prevent abuse of the server.

Usage

We have been monitoring usage of the ORC and have found users accessing the ORC from around the globe, primarily from Europe and the US. So far, between August 31st 2024 and 9th April 2025 209,053 records have been processed from a total of 32 unique emails.

Future Development Goals

Fuzzy Matching

Instead of matching by DOI, we intend to implement a system which matches by Title, Author, Year and other identifying information, including a fuzzy matching step to allow for small differences in metadata, such as abbreviated names. This may be implemented alongside or directly as an optional BibTeX input.

Reference Lookups

Alongside the direct DOI to WorkID conversion, a feature allowing lookup of available references in OpenAlex for all papers identified is planned.

Reverse Lookups

Reversing the ORC to allow for libraries to identify which sources in OpenAlex are also in proprietary bibliometric databases via WorkID to DOI conversion has been requested and is in process of being implemented.

Network Visualisation

A planned extension of the ORC includes allowing for lightweight bibliometric analysis, transforming the ORC into an analysis platform. This includes incorporating a Neo4J instance into the codebase to allow for a visualisation of an OpenAlex dataset in the form of a graph.

Conclusion

The ORC enables bulk conversion of DOIs to OpenAlex WorkIDs, and allows for the generation of sharable research datasets,

[4] github.com/jhculb/Open-Research-Converter

increasing the reproducibility and openness of bibliometric research. It is being utilised by the Scientometrics community, and following user feedback is being expanded into an open-source, lightweight analysis platform for bibliometric analyses.

Acknowledgments

This work was funded by the Federal Ministry of Education and Research via funding numbers: 16WIK2301B / 16WIK2301E, The OpenBib project (Schmidt et al., 2024).

The authors thank Ahsan Shahid for building the web interface of the ORC and Nina Smirnova for providing the problem that inspired the ORC.

References

Culbert, J. H. (2024, November 26). *The Open Research Converter*. https://doi.org/10.5281/zenodo.14222479

Culbert, J. H., Shahid, M. A., & Mayr, P. (2024). *ORC: The Open Research Converter*. https://orc-demo.gesis.org/paper

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). *Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus* (arXiv:2401.16359). arXiv. https://doi.org/10.48550/arXiv.2401.16359

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (arXiv:2205.01833). arXiv. https://doi.org/10.48550/arXiv.2205.01833

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., Taubert, N., Bausenwein, T., & Stahlschmidt, S. (2024). *The Data Infrastructure of the German Kompetenznetzwerk Bibliometrie: An Enabling Intermediary between Raw Data and Analysis*. Zenodo. https://doi.org/10.5281/zenodo.13932928

Velden, T., Hinze, S., Scharnhorst, A., Schneider, J. W., & Waltman, L. (2018). Exploration of reproducibility issues in scientometric research. *STI 2018 Conference Proceedings*, 612–624.

Vieira, G. A., & Leta, J. (2024). biblioverlap: An R package for document matching across bibliographic datasets. *Scientometrics*, *129*(7), 4513–4527. https://doi.org/10.1007/s11192-024-05065-5

https://doi.org/10.51408/issi2025_200

Can Large Language Models Accurately Discriminate Subject Term Hierarchical Relationship?

Yuanxun Li[1], Hongyu Wang[2], Kaiwen Shi[3], Xiaoguang Wang[4]

[1]338018@whut.edu.cn, [2]hongyuwang@whut.edu.cn
School of Management, Wuhan University of Technology, Wenzhi Street, Hongshan District, Wuhan (China)

[3]shikaiwen@whu.edu.cn, [4]wxguang@whu.edu.cn
School of Information Management, Wuhan University, Bayi Street, Wuchang District, Wuhan (China)

Introduction

In the fields of scientometrics and informetrics, accurately determining the hierarchical relationships between subject term is crucial for literature retrieval, domain ontology modeling, and knowledge graph construction. The series of Klink algorithms proposed by Osborne infer relationships between research keywords by integrating multiple data sources and using co-occurrence analysis (Osborne, F., & Motta, E.2012). However, these algorithms struggle with issues such as high computational complexity and low recall when dealing with large-scale data and complex semantic relationships. To address large-scale literature data, most studies adopt a "recall-discrimination" two-stage approach for determining hierarchical relationships.

With the rise of large language models (LLMs), Xu explored the application of LLMs in complex natural language reasoning tasks (Xu, F., Hao, Q., et al., 2025). Researchers have attempted to use text information to identify semantic relationships between words, while Wang utilized LLMs for named entity recognition and semantic relationship extraction (Wang, Z., Huiru Chen, et al., 2025).

The purpose of this study is to explore whether large language models can accurately determine the hierarchical relationship of subject terms, and to compare the performance of large language models in two different phases mentioned before, so as to derive a framework for the application of large language models in the hierarchical relationship determination task. The code is available on GitHub[1].

Methodology

To assess the accuracy of various discriminative strategies in identifying hierarchical relationships of topic words, the study proposes a two-phase framework consisting of a "recall" phase for generating candidate topic word pairs and a "discrimination" phase for evaluating hierarchical relationships. The study introduces two authoritative knowledge systems as gold standards: OpenAlex Concepts[2] and Computer Science Ontology[3] (CSO 3.4.1, containing 165,913 pairs of hierarchical relationships), as shown in Figure 1.

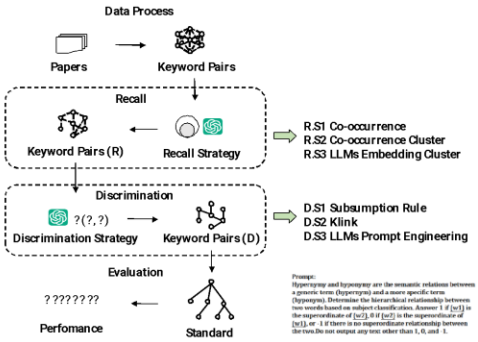


Figure 1. Framework for Discriminating Subject Term Hierarchical Relationships.

[1] https://github.com/Hipkevin/HierarchicalInfer
[2] https://api.openalex.org/concepts
[3] https://cso.kmi.open.ac.uk/downloads

Data Process

This study retrieved literature from the Web of Science Core Collection in the field of computer science from 2010 to 2023 (WC = "Computer Science"), totaling 932,210 articles. Author keywords were extracted from the dataset, and camel case nomenclature was applied to each keyword to ensure its semantic integrity (e.g., "information science" was transformed into "InformationScience"). The processed keywords served as potential subject term candidates, providing the foundation for the data used in the "recall" phase.

Subject Terms Recall

R.S1 Co-occurrence: Count the co-occurrence relationships in author keywords, and when the frequency of keyword co-occurrence pairs is greater than the retrieval year interval (14 years), the two keywords in the keyword co-occurrence pairs are used as candidate subject term pairs.

R.S2 Co-occurrence Cluster: Construct the co-occurrence frequency matrix of author keywords on the basis of R.S1, use this matrix to perform K-Means clustering, select the K value corresponding to the change point of the sum of the squared errors (SSE) curve as the number of clusters according to the principle of the elbow method, and then arrange and combine the keywords in each cluster according to the C_N^2 permutation and take them as candidate subject term pairs.

R.S3 LLMs Embedding Cluster: Also based on R.S1, the embedding vectors of the author keywords are obtained by using a LLM with a smaller number of parameters after distillation, and the candidate subject term pairs are obtained according to the clustering and permutation methods in R.S2.

Hierarchical Relationship Discrimination

D.S1 Subsumption Rule: For keywords x and y of a candidate subject term pair, $P(x|y)$ and $P(y|x)$ are computed, and x is the hypernymy of y when $P(x|y) = 1$ and $P(y|x) < 1$. Usually, the condition $P(x|y) = 1$ is relaxed to $P(x|y) > \alpha$, and α is chosen according to different domains and data sizes, usually 0.8.

D.S2 Klink: Semantic features are introduced on the basis of D.S1 to compute $L(x,y) = (P(x|y) - P(y|x)) * c(x,y) * (1 + N(x,y))$. $c(x,y)$ denotes the cosine similarity of keywords x and y in the co-occurrence matrix, and $N(x,y)$ denotes the string similarity of keywords x and y. In this study, the longest common subsequence distance (LCS) is used. x is the hypernymy of y for $L(x,y) > t$, and t is usually taken as 0.2.

D.S3 LLMs Prompt Engineering: The hierarchical relationship of each candidate subject term pair is discriminated by prompt engineering. The prompt template designed[4] in this study is as follows: `Hypernymy and hyponymy are the semantic relations between a generic term (hypernym) and a more specific term (hyponym). Determine the hierarchical relationship between two words based on subject classification. Answer 1 if {w1} is the superordinate of {w2}, 0 if {w2} is the superordinate of {w1}, or -1 if there is no superordinate relationship between the two.Do not output any text other than 1, 0, and -1`.

Result and Discussion

The accuracy of the recall and discrimination strategies in the experiments of this study on two datasets is shown in Table 1.

Table 1. The Accuracy (%) of the Recall and Discrimination Phases.

Strategy	OpenAlex	CSO
R.S1	**33.51**	**33.22**
R.S2	2.58	2.29
R.S3 (32b)	4.61	3.19
D.S1	4.05	3.45
D.S2	24.29	25.68
D.S3 (72b)	**51.42**	**42.49**
D.S3 (32b)	**49.39**	**39.34**

Without the recall strategy, the computational complexity is $\boldsymbol{O(N^2)}$, while with the recall strategy, the complexity of R.S1 is $\boldsymbol{O(N-M)}$ and R.S2 and R.S3 are $\boldsymbol{O(logN)}$. R.S1 based on co-occurrence frequency truncation performs best, while the LLM embedding-based clustering recall method outperforms the co-occurrence matrix clustering method.

[4]https://en.wikipedia.org/wiki/Hypernymy_and_hyponymy

In the discrimination strategy, D.S3 of qwen2.5 with 72b and 32b parameters correctly identifies all candidate subject terms recalled by R.S1, significantly outperforming the traditional co-occurrence analysis-based discrimination methods. Overall, both co-occurrence analysis and word embedding can detect hierarchical relationships from a semantic perspective. The co-occurrence relationships in R.S1 are broader, while the LLM word embedding provide more precise hierarchical information.

Conclusion

In this study, we propose a framework for the application of LLM in the subject term hierarchical relationship determination according to the two-stage approach of "recall-discrimination", and empirically demonstrate it on large-scale literature datasets in the computer science field. The results show that the large language models can accurately determine the hierarchical relationship between subject terms by relying on the zero-shot capability alone, and an efficient and accurate recall strategy is needed to realize the application framework on large-scale datasets. Follow-up studies can be carried out to optimize the clustering recall method.

Acknowledgments

This work was funded by the National Social Science Fund of China (21&ZD334) and National Natural Science Fund of China (No. 72404215).

References

Osborne, F., & Motta, E. (2012). Mining semantic relations between research areas. In The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11 (pp. 410-426). Springer Berlin Heidelberg.

Wang, Z., Chen, H., Xu, G., & Ren, M. (2025). A novel large-language-model-driven framework for named entity recognition. *Information Processing & Management*, *62*(3), 104054.

Huang, S., Huang, Y., Liu, Y., Luo, Z., & Lu, W. (2025). Are large language models qualified reviewers in originality evaluation? *Information Processing & Management*, 62(3), 103973.

https://doi.org/10.51408/issi2025_188

Delayed Recognition of Novel Ideas: Initially Underestimated, Ultimately Rewarded

Tao Zhiyu[1], Liu Xiaoping,[2], Liang Shuang[3], Li Hanxi[4]

[1]*taozhiyu@mail.las.ac.cn*, [2]*liuxp@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[3]*liangs1998@126.com*
Department of Information Management, Peking University, Beijing (China)

[4]*lihanxi24@mails.ucas.ac.cn*
School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

Introduction

Groundbreaking research challenging established paradigms often sparks debate, yet its growth dynamics and long-term impact are underexplored. Using SciSciNet data, we measure research novelty with the atypical combination index, assess long-term impact via the WSB model, and gauge recognition time with the beauty coefficient. Applying OLS regression, we analyze how innovation strategies influence recognition speed and impact. Findings show that bolder knowledge combinations take longer to gain recognition but yield greater impact. Paradigm-shifting innovations require the longest recognition time but achieve the highest impact, highlighting the rewards of high-risk research. The results revealed by this study have rich implications for innovation polices.

Research Design

To investigate the recognition timeline and impact of novel ideas, we employ the beauty coefficient (SB_B) introduced by Ke et al. (2015), to assess the duration required for a study to gain recognition. Additionally, we apply the WSB model (Wang et al., 2013) to evaluate the ultimate impact of research and the atypical combination indicator (AC) (Uzzi et al., 2013) to quantify research novelty. By incorporating control variables that may affect research impact and using ordinary least squares (OLS) regression, we quantitatively analyze the relationships between dependent and independent variables to elucidate the growth dynamics and long-term impact of novel idea.

The Dependent Variables

We introduce two dependent variables:

(1) The SB_B index, which quantifies the time required for a study to gain peer recognition, defined as follows:

$$SB_B_i = \sum_{t=0}^{t_m} \frac{\frac{c_{t_m} - c_0}{t_m} \cdot t + c_0 - c_t}{\max\{1,\ c_t\}} \quad (1)$$

In this equation, B = 0 for papers with $t_m = 0$. Papers with citations growing linearly with time ($c_t = \ell_t$) have B =0, B is the nonpositive for papers whose citation trajectory c_t is a concave function of time.

(2) WSB model is introduced to quantify the ultimate impact one paper may have in the future, which enable papers to move beyond current impact, which can be defined as:

$$UI_i^{\infty} = m\left(e^{\lambda_i} - 1\right) \quad (2)$$

Where the UI predicts that the total number of citations acquired by a paper during its lifetime, which depends on the relative fitness λ of each paper.

The Independent Variables

Our study investigates how novel ideas shape their ultimate impact and recognition timeline. We utilize the atypical combination (AC) indicator to assess idea novelty and examine its influence on recognition time and impact. The AC is calculated as follows:

$$z = \frac{obs - \mu}{\sigma} \quad (3)$$

Where obs is the observed frequency of the journal pair in the actual WOS, while the μ is the mean and σ indicate the standard deviation. Frequently, the 10^{th} percentile and median z-score are used to describe the novelty of paper from different perspectives.

The Control Variables

Drawing on prior research, we identify variables that simultaneously affect ultimate impact and recognition time, potentially introducing endogeneity with our variable. We include citation count (CC), reference count (RC), funding (FD), atypical combination pairs (AP), and team size (TS) as control variables, with publication year as a fixed effect.

The Evaluation Model

Based on the previous analysis, we design the regression model to implement further evaluation:

$$DV_i = \alpha_i + \beta_1 * AC_10_i + \beta_2 * AC_{M_i} + \beta_4 * CONTROL_i + Y_t + F_i + \varepsilon_i \quad (4)$$

In this equation, DV_i is the dependent variables employed in this study, Y_t is the fixed effect for the publication year, F_i is the fixed effect of the discipline, and the ε_i indicates the random error.

Empirical Study

Data Source and Preprocessing

We leverage the comprehensive SciSciNet dataset, spanning all disciplines. However, due to challenges in data quality, disciplinary heterogeneity, document-type variability, and the need for sufficient citation history, not all data are suitable for analysis. We thus focus on physics, a well-studied field, analyzing 733,648 records from 1892 to 2011.

Main Results

Novel ideas are more prone to delayed recognition (Fig. 1). Analysis of Fig. 1 reveals that 59.15% of highly delayed papers are novel, compared to only 38.44% of instantly recognized papers. This suggests novel ideas face greater delays in recognition, whereas conventional studies are more likely to gain immediate recognition.

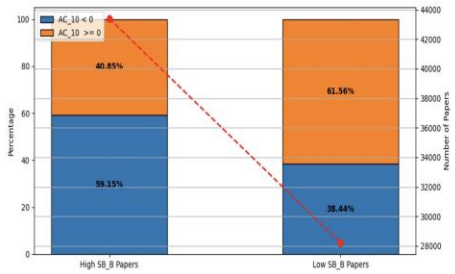


Figure 1. The distribution of SB_B and novelty.

Novel ideas tend to achieve greater ultimate impact. Figure 2 shows that novel ideas have higher impact, with moderately novel research (AC_7) exhibiting a greater likelihood of high impact compared to the most radical ideas (AC_1 and AC_2).

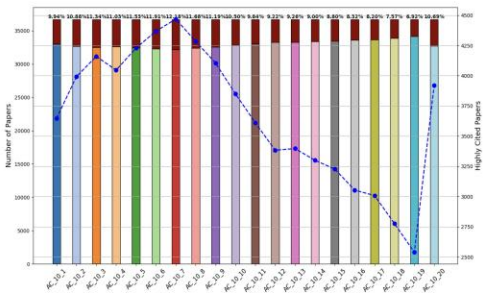


Figure 2. The distribution of novelty and UI.

Research with high ultimate impact is more likely to experience delayed recognition. Figure 3 shows that 33.42% of highly cited papers face significant delays in recognition, whereas low-cited papers rarely exhibit delayed recognition, suggesting that impactful ideas may face initial challenges but ultimately gain substantial recognition.

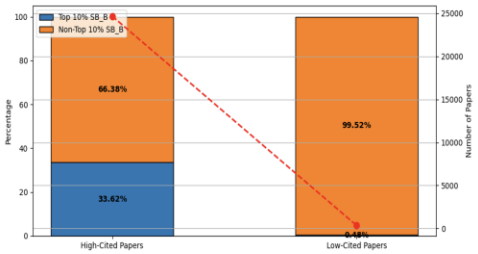


Figure 3. The distribution of SB_B and impact.

We classify innovations into four quadrants based on AC_10 and AC_M. Figure 4 reveals that paradigm-shifting ideas (top 5% AC_M, not top 5% AC_10) exhibit significantly greater ultimate impact than other innovation types but are most likely to experience delayed recognition. This suggests that high-impact ideas, which propose novel explanations for existing problems, often challenge established paradigms, facing resistance from mainstream adopters and resulting in prolonged recognition timelines.

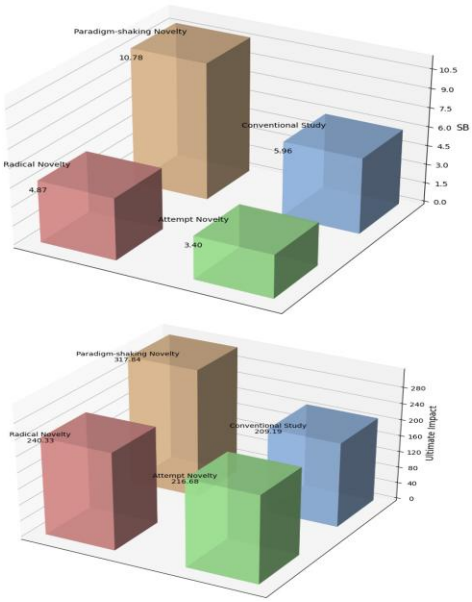


Figure 4. Four types of research and its SB and UI.

Conclusion

By analyzing extensive publication data, our study reveals: (1) Novel research typically experiences longer recognition times and is more likely to be a "sleeping beauty" than conventional research. (2) Moderately novel ideas have the highest probability of achieving high impact compared to highly radical or conventional research. (3) Most high-impact research faces delayed recognition, suggesting that the path to success can be challenging. Further, we derive the following insights: (1) Research evaluation should extend beyond immediate impact to consider long-term scientific contributions. (2) The academic community should foster greater inclusivity toward novel, paradigm-shifting ideas that challenge mainstream theories, as breakthroughs often stem from bold innovations.

References

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proc Natl Acad Sci U S A*, *112*(24), 7426-7431.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468-472.

Wang, D., Song, C., & Barabási, A.-L. s. (2013). Quantifying Long-Term Scientific Impact. *Science*, *342*(6154), 127-132.

https://doi.org/10.51408/issi2025_177

Does winning an Ig-Nobel Prize have an impact on the visibility of the winners' research work?

Philippe GORRY

Philippe.gorry@u-bordeaux.fr
Bordeaux School of Economics CNRS UMR 6060, University of Bordeaux, 16 ave Leon Duguit, 33608 Pessac (France)

Introduction

The Ig-Nobel Prize is an annual award that celebrates unusual or insignificant achievements in the field of scientific research. Created in 1991 by Marc Abrahams, editor-in-chief of the science humour magazine 'Annals of Improbable Research', the Ig-Nobel Prize aims to reward research that 'first makes people laugh, then makes them think'. The prizes are awarded at a humorous ceremony at Harvard University, parodying the more serious Nobel Prizes. Despite their humorous nature, the Ig Nobel prizes often highlight genuine scientific research that is innovative, thought-provoking or simply entertaining. They are a reminder that science can be both serious and fun, and encourage curiosity and creativity in research. The prize committee is examining two sociologically different cases: the first concerns research that could not be reproduced, which is relatively rare among the nominees. This type of research is often already criticised and ostracised in scientific circles as being unscientific, and the Ig Nobel uses humour as a euphemism to denounce and call to order deviant scientists. The second case concerns research that 'should not be reproduced', which brings to public attention peer-reviewed scientific results that would not normally be covered by the media. This situation is more delicate, as it could be perceived by scientists as criticism of legitimate work.

The interpretation of the Ig-Nobel event varies according to one's position in the scientific field. Despite their comical nature, the prizes often highlight genuine scientific research that is innovative, thought-provoking or simply entertaining. They are a reminder that science can be both serious and fun (Gingras & Vecrin, 2002).

To date, the prize has not been the subject of an in-depth bibliometric study, with the exception of Andy Yeung (2022). He analysed 89 articles by prize-winners between 2011 and 2020, and found an average of 42.5 citations per article, with an impact factor of 3.476. It also measured their impact on social networks: 947.3 mentions on Facebook and 263.2 mentions on Twitter. Half of the articles were published in leading journals, and the winners were recognised within 2 years.

We decided to revisit this work and examine whether winning the Ig Nobel Prize had an impact on the visibility of the scientists' work concerned, by analysing changes over time in the citations of the articles concerned.

If the Ig-Nobel Prize is perceived as a criticism, the research articles concerned should record a decrease in citations If the prize is perceived with humour, it may attract the attention of the scientific community and lead to an increase in the citations of the references cited. Alternatively, the scientific community is insensitive to the Ig-Nobel prize, and citations of cited work, as well as the reputation of scientists, are unaffected.

We have taken into account all the Ig-Nobel prizes since 1991, for which a scientific article is referenced on the 'Improbable Research' site, whatever the prize-winning field. However, we only included those for which the scientific reference was registered in the Scopus database. We then carried out a bibliometric analysis of this sample. In a second step, we extracted the scientometric characteristics of the authors, and in a third step, we propose an analysis of both the citation trends of these articles, with a comparison with a matched sample in order to carry out an analysis using the difference-in-difference method.

Methods

The 'Improbable research' website lists 231 Ig-Nobel prizes between 1991 and 2024, covering 54 different disciplines. The top 3 disciplines (with their associated sub-fields) are, in order, medicine (25.6%), biology (19.48%) and physics (19.04%). Other areas of the social sciences (art, literature, management, psychology, etc.) are also represented, as is the Ig-Nobel Peace Prize. In all, 166 annual prizes in each discipline have at least one associated bibliographic reference (71.86%). But out of a total of 231 references, only 180 have been identified in the Scopus database with a citation history.

This initial sample was then subjected to a bibliometric analysis (source, year, citations) with the extraction of temporal citation data. From this initial sample, a second panel was created with all authors for a second bibliometric analysis (affiliation, country, number of publications, citations and co-authors, and h-index). Thirdly, the sample of publications was matched with two control populations, one composed of articles published in the same year in the same journal, the other composed of articles matched by keywords and published in the same year in the same journal.

Statistical correlation tests were performed to compare the number of citations before and after the Ig Nobel Prize, and to compare these differences between the Ig Nobel Prize publication sample and the control samples using the difference-in-differences method using the STATA software package (Villa, 2016). Only articles published at least 3 years before the authors received the Ig-Nobel Prize were included in the statistical analyses, a total of 86 articles.

Preliminary Results

Bibliometric analysis of Ig-Nobel Prize publications

A sample of 180 articles published between 1967 and 2020 was collected from Ig-Nobel prizes awarded during the period 1991-2020 (data not shown). Most of the articles are of the article type (89.44%), but all other types are represented. They were published in 133 different scientific journals, with 57.89% belonging to the first quartile, 23.31% to the second quartile, 9.77% to the third and only 2.26% to the fourth quartile. 90.6% of the articles had fewer than 241 citations at the end of 2024, around an average of 148.4 and a median of 32 citations, with a maximum of 4,704 citations.

Bibliometric analysis of the Ig-Nobel prize winners

The first sample enabled us to characterise a population of 234 different authors. It is made up of 31 different nationalities: the top 5 nationalities are, in order: US American (21.37%), English (13.31%), Japanese (12.10%), French (7.66%) and Dutch (6.05%). The authors are affiliated with 162 different institutions, including 35 universities ranked in the top 100 of the Academic Ranking of World Universities, including Harvard University and Stanford University. In median terms, the authors awarded the Ig-Nobel Prize published 54 articles, with 103 co-authors, accumulated 1,596 citations and had an H-index of 22 (table 1).

Table 1. Bibliometric characteristics of authors.

Statistics	Citations	Documents	h-index	Co-authors
Nb.	234	234	234	234
Min.	1	1	1	0
Max.	302612	2201	195	8039
1st Quartile	264.75	11.25	7.25	16.25
Median	1596,00	54,00	22,00	103.5
3rd Quartile	5547,00	150,00	39,00	293.5
Mean	9224.39	131.89	30.09	351.78
SDT	27350.08	226.53	31.98	835.78

However, there are wide variations across all indicators, with extremely high maximum values (2201 publications, 302612 citations and an H-index of 195). All variables have a highly skewed distribution to the right (data not shown). A similarity matrix highlights the correlations between all the variables, with the maximum value observed between the H-index and the number of documents published (data not shown).

Statistical test of correlation of citations before/after the Ig-Nobel prize

If we compare the number of annual citations for each article in the three years preceding the award of the Ig Nobel Prize with the citations in the three years following, we find a

persistent and significant increase in the number of annual citations. A comparison of the variance in citations of publications before and after the award of the Ig Nobel Prize reveals a p-value of less than 0.006. As the calculated p-value is below the significance level alpha=0.05, we must reject the null hypothesis H0 and retain the alternative hypothesis according to which there is indeed a significant difference between the number of citations in the years before and after the award of the Ig Nobel Prize.

Discussion & Perspectives

With a sample size twice as large and spread over a period twice as long as that of Yeung (2022), our bibliometric analysis of the bibliographic references of the Ig Nobel Prize winners is fairly similar. Most of the articles were published in top-quartile journals and received a large number of citations.

The bibliometric characteristics of the prize-winners have been analysed for the first time, revealing productive researchers who collaborate widely and whose work is recognised. Almost a third of them belong to internationally ranked universities, some of them very prestigious (Harvard, Stanford, Oxford). Finally, the preliminary results show that the number of citations received per article is not negatively affected. There is even a significant increase in the number of annual citations after the Ig-Nobel prize is awarded.

However, these initial results remain to be confirmed using the difference-in-difference method, and using matched publications as the control population.

References

Gingras, Y. & Vécrin, L. (2002) Les prix Ig-Nobel. In : *Actes de la recherche en sciences sociales* (pp. 66-71). Paris: Persée.

Villa, J.M. (2016). Diff: Simplifying the estimation of difference-in-differences treatment effects". *The Stata Journal*. 16, 52–71.

Yeung, A. (2022). Not just nickel-and-dime: An analysis of journal articles winning Ig Nobel prize during 2011–2020, *Malaysian Journal of Library & Information Science*, 27, 93-99.

https://doi.org/10.51408/issi2025_169

Enlarging the spectrum. Implementing a local extension of ROR as identification instrument for additional actors in Flemish (SSH) research

Peter Aspeslagh

peter.aspeslagh@uantwerpen.be
Centre for R&D Monitoring, University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract

International organization databases, like the Research Organization Registry (ROR), are essential tools to identify unique organizations for a wide range of purposes. Our experience with the structural enrichment of author affiliation data for SSH publications pointed to opportunities that a local extension of an international database can provide. Applying a step-by-step approach, we are steadily extending the original set of Flemish/Belgian ROR organizations, building a dataset that contains unique organization identifiers covering a broader spectrum of research, educational and governmental organizations, linking multiple datasets in a national environment. In this poster, we will present the context and methodology related to the creation of this new instrument, the Flemish Organization Registry.

Introduction

In 2019, a parameter measuring international collaboration was added to the Flemish performance-based research funding system (PRFS). Therefore, for SSH disicplines, author affiliation data had to be provided for the full set of publications included in the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB). This dataset is a compilation of approved publications authored by all researchers affiliated to a SSH faculty at a Flemish university. However, author affiliation data was not available in VABB.

As only less than half of the VABB publications are included in the Web of Science (Aspeslagh & Guns, 2023), we launched a large and multifaceted author affiliation data collection operation. Browsing alternative databases like Scopus or OpenAlex only resulted in data for a limited percentage of publications; most of the cases required manual intervention for the retrieval of the actual affiliation data.

During the subsequent coding we assigned unique (ROR-)identifiers to each author affiliation. However, the coding process showed that only two thirds of the affiliated organizations were covered by ROR. In order to allow the completion of the project – and to comply with the new parameter in the Flemish PRFS – new organizations were added to a local copy of ROR and coded according to the ROR data scheme. For the 2012 to 2022 time window, 34.7% of the unique organizations found in author affiliation data of the non-Web of Science publications did not have an identifier in ROR (n=2,348).

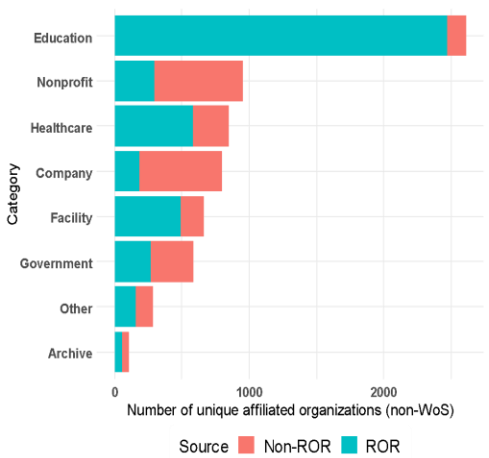


Figure 1. Number of unique affiliated organizations for non-WoS publications (ROR vs non-ROR organizations).

When taking into account the (ROR) type variable as categorization, ROR very much covers the Education category (Figure 1), but

less so for Company, Nonprofit and Government.

National extension

The addition of non-ROR organizations for a complete coverage of author affiliation data demonstrated that there is room for a national extension of an international database. Such an expansion provided a new opportunity. The novel unique identifiers and related metadata could be valorised in a broader context: not only in function of a PRFS parameter, but also as tool to map the full spectrum of organizations that are involved in research in Flanders, independent of their inclusion in common international databases. Too often, organization data is scattered over different lists with distinct identifiers: lists with organizations only having a EU Participant Identification Code (PIC), selections of the entities included in the Belgian commercial register (KBO) etc. With the development of a new instrument, the Flemish Organization Registry (FOR), an integration of different lists is envisaged.

Compilation

Due to the recurrent author affiliation data collection, a platform for hosting, managing and enriching organization data was already available. This allowed us to shift from a technical to a content-related perspective, focussing on the addition of an extensive set of Flemish/Belgian organizations not included in the current dataset.

During a first phase, we made an inventory of the different commonly used organization lists in a broad research, governmental and educational context in Flanders. Deployed for diverse purposes, these lists often contained equal organizations but were not (entirely) interoperable.

Secondly, the selected new datasets, as well as the extended ROR database, were consolidated both by matching via the available metadata or by manual intervention. Often, this completed types of organizations of which our original database contained some, but not all. For example, 20 municipal administrations, found as author affiliations during the affiliation data collection project, were added as these organizations (local governments) were not available in ROR. The consolidation phase provided the remaining 561 Belgian municipal administrations. When future registering of an author affiliated to the 21st municipal administration, the organization will be available in the database. In a third step, which is ongoing, we are uploading the newly consolidated organizations to the main FOR database. Relevant metadata that was available in the original lists is also being added to the database, which is compatible with the ROR data model. Often, this results in a single organization ID now containing a PIC, KBO, FOR and, if available, a ROR identifier. It will allow the continued addition of metadata and enabling, among other features, customized categorization.

Governance and future developments

As FOR can serve diverse purposes in multiple entities in the Flemish governmental and research landscape, a governance framework is being established to ensure the extended use of this project.

Even if FOR places an emphasis on Flemish/Belgian organizations, the relevant part of the new additions and related metadata can be transferred to international databases.

The database allows the continued addition of new variables and categories, which can be used to add new dimensions to future bibliometric studies from an organizational perspective.

Conclusion

With this extended organization database, the efforts invested in the implementation of a modification in the Flemish PRFS are being extra valorised. The format created for the registration of additional organizations during the author affiliation data collection is now being developed into an instrument to be used for multiple purposes. The step-by-step approach to compile and consolidate a diverse set of organization lists results in a Flemish Organization Registry, enabling the unique identification of a broader spectrum of actors in Flemish (SSH) research. While the focus is currently on Flemish/Belgian organizations, the data scheme is compatible with ROR with the possibility to transfer relevant data to external databases.

Reference

Aspeslagh, P. & Guns, R. (2023). How international is co-authorship outside the Web of Science? The case of social sciences and humanities in Flanders, Belgium. Proceedings of ISSI 2023 – the 9th International Conference of the International Society for Scientometrics and Informetrics, 2, 29-36. https://doi.org/10.5281/zenodo.8350379

https://doi.org/10.51408/issi2025_202

EU-Armenia Scientific Partnership: A Bibliometric Analysis of Funding and Academic Output

Ruzanna Shushanyan[1], Maria Ohanyan[2], Miranush Kesoyan[3], Mariam Yeghikyan[4], Gevorg Kesoyan[5]

[1]*ruzanna.shushanyan@ysu.am*, [2]*mohanyan226@gmail.com*, [3]*mkesoyan1996@gmail.com*, [4]*mariamyeghikian@gmail.com*, [5]*gevorgkesoyaned@gmail.com*

Center for Scientific Information Analysis and Monitoring, Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia, Yerevan (Armenia)

Introduction

The European Union (EU) plays a crucial role in providing substantial funding for scientific research across its member countries and outside through the Framework Programs for Research and Technological Development (FPs) (Gallo et al., 2021). The EU has established itself as a major contributor to research and innovation through the FPs (Dalen et al., 2024), which have become a central source of funding for both applied and basic research (Enger & Castellacci, 2016), covering cooperation with the EU-neighboring countries, including Armenia. EU-Armenia scientific collaboration is mainly focused on EU-supported projects under FP7 and, more recently, Horizon 2020. This form of 'science diplomacy' serves not only as an instrument of external policy but also as a means to enhance relations between the EU and its partners (Mazepus et al., 2017).

In this study, we examine the scientific output and cooperation between EU-funded major projects (such as the FP7, Horizon, and Horizon2020) and Armenia through bibliometric analysis. We aim to elucidate the year-wise distribution of scientific publications being published within the EU-funded projects, along with the relevant academic fields represented in these publications; and the contribution of Armenia as a partner, highlighting the most active organizations involved in these projects.

Data and methods

Following a structured bibliometric methodology, this paper examines the participation of Armenia in EU-funded scientific research projects within the framework of a new dimension of funding acknowledgment analysis. We conducted a bibliometric analysis by extracting publication metadata from the Web of Science (WOS) database through country-based filtering, funding number identification, and source verification. Whereas traditional bibliometric studies only focus on tracing co-authorship or citation networks, here we incorporate the analysis of funding acknowledgment to directly trace the contribution of EU financial support. Afterward, the harnessed publications were categorized based on their references to EU projects. The retrieved data was analyzed to illustrate the total number of publications indexed in the WOS repository; the academic fields and disciplines were sorted according to Glänzel and Schubert's classification (Glänzel & Schubert, 2003) of subject categories. It's important to note that the full counting method was used when the publication simultaneously has full value for all specified areas.

We also utilized network analysis and techniques to reveal the role of Armenian organizations in the research networks funded by the EU by combining funding acknowledgment analysis with network-based institutional evaluation.

Results and Discussion

The conducted bibliometric analysis revealed that about 2408 publications were published and indexed in the WOS database through the EU-funded projects in the frames of the EU-Armenia partnership. The highest number of publications was recorded in the years 2016 and 2018. However, following 2018, the amount of publications dropped but

has continuously increased since the year 2023.

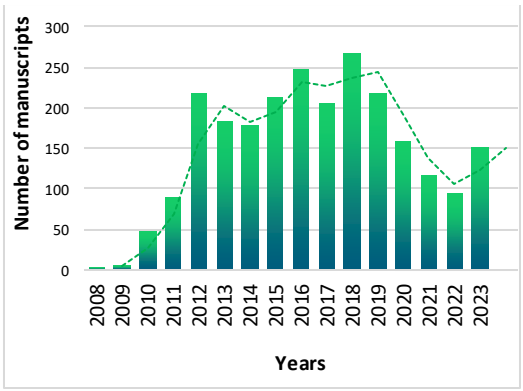


Figure 1. Number of publications published within EU-funded projects by year.

Notably, the fields of Physics and Geosciences & Space along with Multidisciplinary sciences exhibit the highest number of publications, indicating that EU-funded projects mainly focus on applied sciences.

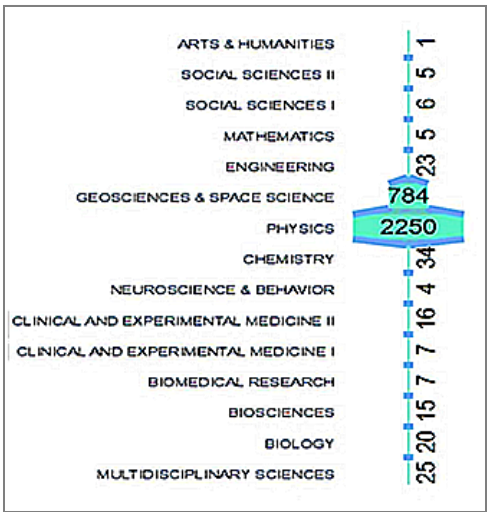


Figure 2. Total number of publications categorized by Glänzel and Schubert classification.

Particularly, the highest number of successful projects occurred in 2019 and in 2023, which aligns with the overall trends in the published publications.

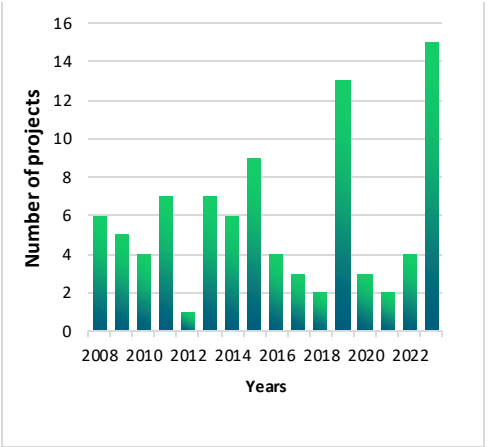


Figure 3. The year-wise distribution of EU-funded projects.

The sharp increase in publications in 2020 and 2023 may be explained by the large-scale projects or special funding initiatives, while fluctuations in other years suggest variations in project approval rates or policy shifts. Based on our analysis, Armenia has played a diverse role in promoting EU-funded projects. In most instances, Armenia acts as a participant (73%), while in some projects, it serves as a coordinator (18%) and a partner (8.1%). In only a few cases, Armenia plays a minor role as a third-party country, accounting for 0.9% of the projects. These results demonstrate that Armenia has engaged in various EU-funded projects in different capacities, which have contributed to an increase in publications and fostered connections among scientific collaborators.

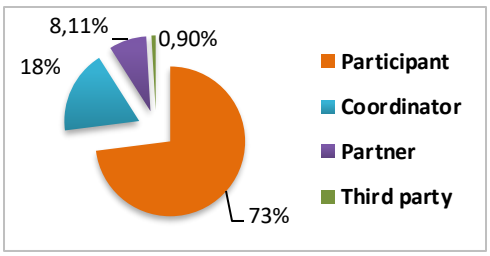


Figure 4. The contribution of Armenia within EU-funded projects.

In total, 29 organizations participated in EU-funded projects during the studied years; however, the most active organization is the National Academy of Science of Armenia.

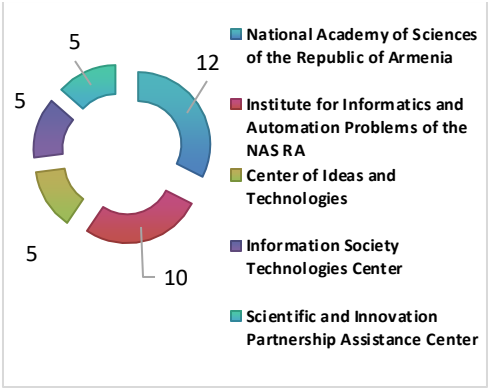


Figure 5. The top 5 active Armenian organizations participating in EU-funded projects.

Figure 5 illustrates the list of these five organizations. It is evident that the majority of the projects focus on the IT sector and related fields, which in turn indicates the priorities and policy direction of the EU in the context of a scientific partnership with Armenia and emphasizes the academic potential of the mentioned areas.

Conclusion

The outcomes of our bibliometric study represented the overall picture of the EU-funded projects, detailing the trends in scientific output, the diversity of academic fields engaged, and the active contributors associated with programs carried out with the support of the EU's financial and collaborative efforts. This investigation provides findings into the nature and extent of Armenia's involvement in EU-funded research initiatives, highlighting both achievements and areas for potential growth. By elucidating the contributions and evidence associated with EU-funded projects, we are keen to inform future policy decisions and foster enhanced cooperation in research and innovation between Armenia and the EU.

Additionally, the findings of this study demonstrate how funding acknowledgment analysis can enhance bibliometric assessments of international research collaborations. By integrating publication output trends with institutional network analysis, we offer a framework that can be adapted to other regions involved in EU-funded projects. Further, we will conduct a comparative analysis of research collaboration between the EU and the South Caucasus region utilizing co-authorship institutional network mapping, and financial output assessment. This could give an insight into the EU scientific policies and funding impact at the regional level.

References

Gallo F., Seniori Costantini A., Puglisi M. et al. (2021). Biomedical and health research: an analysis of country participation and research fields in the EU's Horizon. *Eur J Epidemiol* 36, 1209–1210.

Dalen van Ryanne et al. (2014). Public funding of science: An international comparison. *CPB Netherlands Bureau for Economic Policy Analysis*. 4-5.

Enger, S. & Castellacci, F. (2021). Who gets Horizon 2020 research grants? Propensity to apply and probability to succeed in a two-step analysis. *Scientometrics* 109, 1611–1612.

Mazepus H., Toshkov D., Ramasheuskaya I., Chulitskaya T., Rabava N. (2017). *The Effects of the EU's Scientific Cooperation Programmes on the Eastern Partnership Countries: Scientific Output and Broader Societal Impact*. 35.

Glänzel W., Schubert A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics* 56, 357–367.

https://doi.org/10.51408/issi2025_182

Evaluating Large Language Models for Gender Bias in Academic Knowledge Production

Judit Hermán[1], Kíra Diána Kovács[2], Yajie Wang[3], Orsolya Vásárhelyi[4]

[1]*hermanjudit01@gmail.com*, [2]*kira20020111@gmail.com*
Budapest University of Technology and Economics, Faculty of Natural Sciences, Budapest, Hungary

[3]*yajie.wang998@gmail.com*
Center for Collective Learning, Corvinus Institute for Advanced Studies, Corvinus University, Budapest, Hungary

[4]*orsolya.vasarhelyi@uni-corvinus.hu*
Center for Collective Learning, Corvinus Institute for Advanced Studies, Corvinus University; Institute of Data Analytics and Information Systems, Corvinus University, Budapest, Hungary

Introduction

Gender inequality persists in science, with women being underrepresented in leadership and disadvantaged in hiring, funding, and publishing. While Large Language Models (LLMs) like ChatGPT and Gemini offer new tools for research support, they also risk reinforcing existing biases. Prior studies show LLMs can reproduce gender and racial stereotypes, hallucinate references, and generate inconsistent outputs. This study evaluates references produced by nine advanced LLMs across 26 research subfields and four major domains, comparing them to the OpenAlex database to assess accuracy, gender balance, publication trends, and consistency.

Related Work

Women remain underrepresented in senior academic roles, especially in STEM, due to barriers like unequal access to resources, limited mentorship, and work-life conflicts (Legewie & DiPrete, 2014; Winslow, 2010; Vásárhelyi, 2020; Hopkins et al., 2013; Huang et al., 2020). LLMs may seem promising for reducing inequalities by equally representing the work of men and women, but they may actually worsen these disparities by reproducing gender and racial biases present in their training data (Ferrara, 2023; Smith & Rustagi, 2021; Zhou et al., 2024; Ghosh & Caliskan, 2023). They also hallucinate references (Metze et al., 2024; Buchanan et al., 2023) and overcite highly cited, male-authored works (Algaba et al., 2024; Antu et al., 2023), potentially reinforcing existing inequalities. Ensuring equity requires critically evaluating AI outputs (Zimmermann et al., 2024; Kotek et al., 2023; Pfohl et al., 2024). Based on these findings, we hypothesized that LLMs undercut women's work.

Data

We analyzed outputs from nine LLMs and used the OpenAlex database of 250+ million publications. From OpenAlex's classification, we selected the 20 most-published topics in 26 subfields across four disciplines, yielding 497 topics. To reduce the size of our data, we included only articles that were cited at least twice within our OpenAlex baseline database for these topics.

Methods

We prompted each LLM with a standardized query to generate literature reviews and references. Hallucinated references were detected using fuzzy string matching based on Levenshtein distance and a Jaccard index filter, with a threshold of 0.86. We inferred authors' gender using a name-based gender and ethnicity inference method, Ethnea (Torvik & Agarwal, 2016). For each paper in both the OpenAlex dataset and the LLM-

generated outputs, we calculated the ratio of female authors. We then analyzed these ratios by averaging the proportion of women at both the subfield and major academic domain levels. Statistical differences were tested using the Mann-Whitney U test ($\alpha = 0.05$) on female authorship and reference matching rates.

Results

Some LLMs (e.g., Claude 3.5 Sonnet, ChatGPT 4o) slightly overcite women, while others (e.g., Gemini models, Llama 3.3 70b, DeepSeek R1) tend to undercite them—often significantly. Citation patterns varied by field: Gemini 2.0 Pro and Llama 3.3 70b cited more women in Health Sciences, while other Gemini models less in Social Sciences. Gender bias persisted even when considering only recent publications, especially in Physical and Life Sciences, indicating model-driven citation patterns.

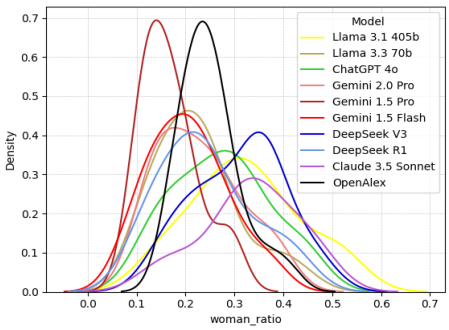


Figure 1. Density of the ratio of women in the OpenAlex and LLMs' references.

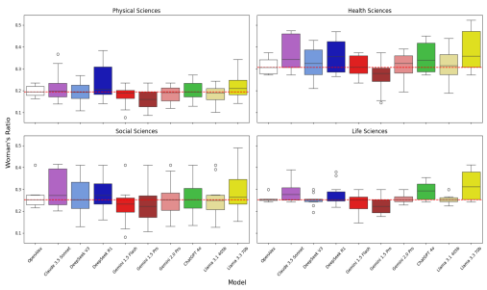


Figure 2. Distribution of female author ratios in LLM and OpenAlex references across four scientific fields. Boxes show quartiles; whiskers indicate non-outlier ranges. Dotted line marks OpenAlex median.

Contrary to earlier findings (Antu et al., 2023), our analysis shows that all examined LLMs now favor recent publications, except Gemini models in Social Sciences. The analysis of moving averages across the years reveals that large language models often show statistically significant differences from OpenAlex in the ratio of women authors across disciplines—especially in Physical Sciences—and these disparities become more pronounced in papers published after 2000, indicating increasingly widespread gender citation gaps.

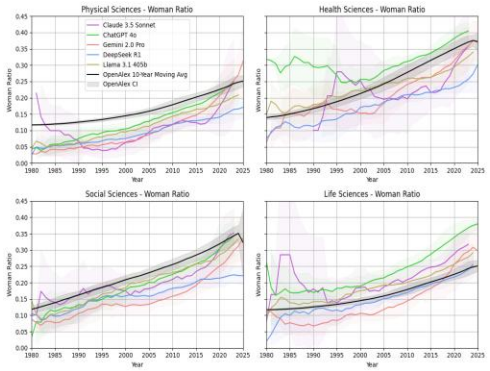


Figure 3. Ratio of women in references in the 4 main fields with moving averages and 95% confidence intervals.

Over 70% of LLM references were hallucinated, with ChatGPT 4o reaching 93% and Gemini 2.0 Pro and DeepSeek R1 the lowest (~70%), underscoring the need for citation caution. Even among real references, models like Gemini 1.5 Flash and Llama 3.1 405b undercited women, while Llama 3.3 70b overcited them—especially in Health and Life Sciences—indicating persistent gender bias.

References

Algaba, A., et al. (2024). Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias.

Antu, S. A., et al. (2023). How ChatGPT Selects Sources: A Study of Bias in Scientific Literature Recommendations.

Buchanan, B., et al. (2023). ChatGPT Hallucinates Non-existent Citations: Evidence from Economics.

Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models.

Ghosh, S., & Caliskan, A. (2023). Stable Diffusion reflects social bias and stereotypes in image generation.
Hopkins, N., et al. (2013). Gender Disparities in Academic Productivity and Representation.
Huang, J., et al. (2020). Historical comparison of gender gaps in scientific publishing and impact.
Kotek, H., et al. (2023). Gender Stereotypes in LLM-generated Texts: An Empirical Investigation.
Legewie, J., & DiPrete, T. A. (2014). The High School Path to STEM: Gendered Influences on Science Orientation.
Metze, K., et al. (2024). Bibliographic Research with ChatGPT may be Misleading: The Problem of Hallucination.
Pfohl, S. R., et al. (2024). Language Bias in AI-generated Reference Letters.
Smith, S., & Rustagi, J. (2021). Gender and Racial Bias in AI Systems: An Audit of 133 Models. Berkeley Haas Center for Equity, Gender, and Leadership.
Torvik, V. I., & Agarwal, S. (2016). Ethnea—an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. Paper presented at International Symposium on Science of Science, Washington DC, United States.
Vásárhelyi, O. (2020). Barriers to Women's Retention in Technology and Engineering.
Vásárhelyi, O., et al. (2021). Gendered Patterns in Online Science Dissemination.
Winslow, S. (2010). Gender Inequality and Time Allocation in Academia.
Zhou, Y., et al. (2024). Stereotypical Visual Representations in AI-generated Images.
Zimmermann, R., et al. (2024). Evaluating ChatGPT's Ability to Generate Literature Reviews.

https://doi.org/10.51408/issi2025_186

Exploring institutional type composition in scientific collaboration and its role in scientific impact

Shuang Liang[1], Zhiyu Tao[2], Qingshan Zhou[3]

[1]liangs@stu.pku.edu.cn, [3]zqs@pku.edu.cn
Department of Information Management, Peking University, Beijing (China)

[2]taozhiyu@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

Introduction

Many significant innovations and advancements in scientific research are achieved through inter-sector collaboration. With the continuous societal development, there is an increasing call for closer collaborative relationships among institutions, aiming to jointly address complex and evolving social challenges as well as technological issues. Interinstitutional collaboration integrates the unique resources and strengths of different institutions, playing a crucial role in improving research performance. It is a strategic approach for academic institutions to enhance funding acquisition and increase academic visibility (Zhou & Tian, 2014). Moreover, it also exerts a significant influence on researchers' academic performance, such as the number of publications and H-index (Bikard, Vakili & Teodoridis, 2019; Zhang & Wang, 2017).

Existing research has largely centered on universities, industries, and government institutions, often overlooking the roles and potential impacts of other institutional types, as well as the broader implications of institutional type composition in scientific innovation. In this study, we undertake a comprehensive examination of the roles of eight institutional types in scientific collaboration and investigate how institutional diversity and power structure influence research performance.

Methods

We obtained information of 26,998 institutions from Sciscinet (Lin et al., 2023). Institution type were further obtained from ROR (Research Organization Registry). Through records matching, all institutions were classified into eight categories: education, company, facility, government, healthcare, nonprofit, archive, and other. We extracted 8,454,850 records of multi-institutional collaborations from 1980 to 2021, each containing institutional information and five-year citation data (C5).

Institutional type diversity is defined as the number of different institutional types involved in academic collaboration. The formula is as follows:

$$D_i = \sum_{k=1}^{K} I_{ik}$$

where D_i denotes the institutional type diversity of paper i, K is the total number of institutional types, which is 8 in this study. If institutional type k is involved in paper i, then $I_{ik} = 1$, otherwise, $I_{ik} = 0$.

We use the Herfindahl Index (HI) to measure the institutional power centralization in academic collaboration:

$$P_i = \sum_{k=1}^{K} \left(\frac{ins_num_{i,k}}{\sum_{j=1}^{K} ins_num_{i,j}} \right)^2$$

where P_i denotes the institutional power centralization for paper i. $ins_num_{i,k}$ represents the number of institutions of type k in paper i. K is the total number of institutional types, with a value of 8. A lower P_i indicates a more balanced distribution of institutional power. In contrast, papers with a high P_i have a higher degree of centralization in institutional power.

Results

Institutional type and scientific impact

We analyzed the number of publications and the 5-year citation performance for different institutional types (Fig.1). Nonprofit institutions demonstrate exceptionally excellent citation performance compared to other types. We created dummy variables for institutional types and controlled for other variables. Taking education institutions as the baseline, we performed a regression analysis to examine the effect of different institutional types on citations (Tab.1).

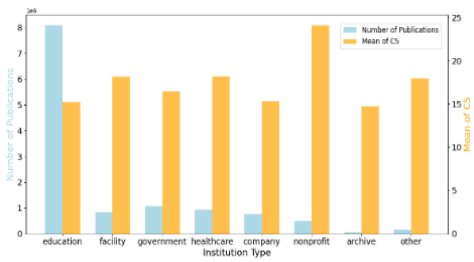


Figure 1. Publication count, C5 performance for different institutional types.

Table 1. Multivariable regression for institutional types and citation impact.

	logC5
company	0.0071** (0.002)
facility	0.1089*** (0.002)
healthcare	0.1119*** (0.004)
government	0.0835*** (0.002)
nonprofit	0.2280*** (0.003)
archive	-0.0469*** (0.005)
other	0.0921*** (0.003)
control	Yes
const	1.1639*** (0.011)
Obs	8454850
R^2	0.211

Note: *** p<0.001, ** p<0.01, * p<0.05, with robust standard errors in parentheses.

Institutional type diversity and scientific impact

We depicted the citation performance of different institutional type diversities (Fig.2) and found that collaboration across more institutional types contributes to higher citation counts. To further investigate whether there is a relationship between the institutional type diversity and citation impact of papers, we conducted the following multivariable regression analysis and reported the results in Table 2.

$$C_i = \alpha + \beta_1(D_i) + \beta_2(D_i^2) + \beta_3(Controls) + \varepsilon_i$$

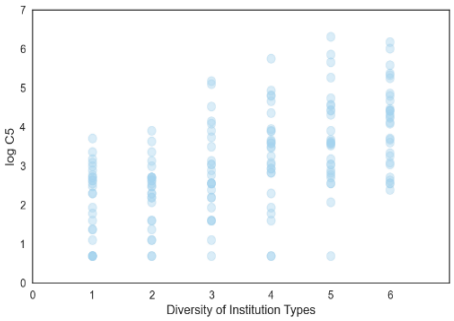


Figure 2. The relationship between institutional type diversity and citation impact.

Table 2. Multivariable regression for institutional type diversity and citation impact.

	logC5	*logC5*
type diversity	0.1168***(0.002)	0.0873***(0.004)
type diversity^2		0.0087***(0.001)
Control	Yes	Yes
const	1.0464***(0.009)	1.0714***(0.011)
Obs	8454840	8454840
R^2	0.210	0.210

Institutional power centralization and scientific impact

We further examined citation patterns across varying levels of institutional power centralization. The results reveal an overall decline in citation impact as power centralization increases (Fig.3). We conducted the following regression analysis to examine the potential nonlinear relationship between institutional power centralization and scientific impact (Tab.3). The estimated turning point occurs at 1.178, beyond the observed range of power centralization. Accordingly, citation impact consistently decreases with increasing centralization of institutional power.

$$C_i = \alpha + \beta_1(P_i) + \beta_2(P_i^2) + \beta_3(Controls) + \varepsilon_i$$

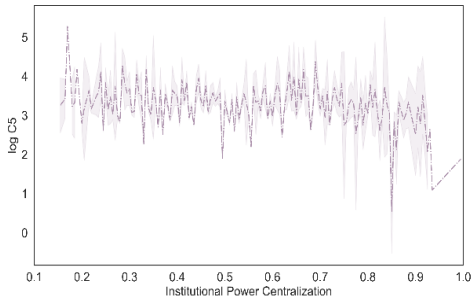


Figure 3. The relationship between institutional power centralization and citation impact.

Table 3. Multivariable regression for institutional power centralization and citation impact.

	logC5	*logC5*
power centralization	-0.2432***(0.004)	-0.6583***(0.028)
power centralization^2		0.2794***(0.018)
Control	Yes	Yes
const	1.3780***(0.014)	1.5139***(0.017)
Obs	8454850	8454850
R^2	0.209	0.209

Conclusion

This study offers a comprehensive categorization of institutional types and identifies statistically significant relationships between institutional type, institutional type diversity, institutional power centralization, and scientific impact. Greater institutional diversity is positively correlated with a significant increase in citation impact. Nevertheless, excessive power centralization in inter-sector institutional collaborations appears to hinder citation performance. The results provide valuable insights for research management and the development of institutional collaboration strategies.

References

Bikard, M., Vakili, K., & Teodoridis, F. (2019). When collaboration bridges institutions: The impact of university-industry collaboration on academic productivity. *Organization Science, 30*(2), 426-445.

Lin, Z., Yin, Y., Liu, L., & Wang, D. (2023). SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data, 10*(1), 315.

Zhang, B., & Wang, X. (2017). Empirical study on influence of university-industry collaboration on research performance and moderating effect of social capital: Evidence from engineering academics in China. *Scientometrics, 113*(1), 257-277.

Zhou, P., & Tian, H. (2014). Funded collaboration research in mathematics in China. *Scientometrics, 99*(3), 695-715.

https://doi.org/10.51408/issi2025_204

Fully Algorithmic Librarian: Large-Scale Citation Experiments

Tomasz Stompor[1], Janina Zittel[2], Thorsten Koch[3], Beate Rusch[4]

[1] *stompor@zib.de*
Zuse Institute Berlin, Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV), Takustraße 7, 14195 Berlin (Germany)

[2] *zittel@zib.de*
Zuse Institute Berlin, Applied Algorithmic Intelligence Department, Takustraße 7, 14195 Berlin (Germany)

[3] *koch@zib.de*
Zuse Institute Berlin, Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV), Takustraße 7, 14195 Berlin (Germany)
Zuse Institute Berlin, Applied Algorithmic Intelligence Department, Takustraße 7, 14195 Berlin (Germany)

[4] *rusch@zib.de*
Zuse Institute Berlin, Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV), Takustraße 7, 14195 Berlin (Germany)

Introduction

Libraries play a crucial role in supporting academic publishing by providing access to bibliometric tools that help researchers navigate vast citation networks (Web of Science, Scopus, OpenAlex). As scientific output grows exponentially (de Solla Price, 1963), algorithmic approaches to citation network analysis are becoming increasingly important. The Fully Algorithmic Librarian (FAN) is an interdisciplinary research project in the fields of mathematics and library and information science, carried out by two departments of the Zuse Institute Berlin. The project's goal is to analyze large-scale citation networks on a knowledge graph sourced from Web of Science and Open Alex citation data that is currently under development. It will serve as a basis for the design of application scenarios. for algorithmic-intelligence-(AI)-supported methods in academic libraries as central research support institutions.

This poster paper aims to present two algorithmic approaches for analyzing large-scale citation networks, which serve as two preliminary steps of the project. Firstly, the results of a comparison of Web of Science (WoS) and OpenAlex databases using the PageRank algorithm reveals key differences. Secondly, a multi-label clustering technique designed for large-scale citation networks accounts for disciplinary variations in publication practices.

A key challenge in bibliometric analysis is the structural and disciplinary diversity of citation networks. Commercial databases like Web of Science (WoS) and open alternatives like OpenAlex (Priem et al., 2022) offer rich but distinct representations of academic publishing. Understanding their differences is essential for developing reliable bibliometric methods. This study demonstrates the effectiveness of algorithmic approaches in analyzing the structural properties of publications in both databases and presents a clustering technique tailored to the varying publication practices across academic fields.

By integrating these insights, this work contributes to the development of automated bibliometric tools that enhance library services, assist researchers in navigating citation landscapes, and support institutions in evaluating academic impact. The findings highlight the potential of algorithmic bibliometry in library and information science and underscore the importance of open, scalable solutions for analyzing scholarly communication.

Comparison of the citation graphs based on WoS and OpenAlex

Evaluating scientific impact requires precise measurement of individual article influence, traditionally assessed through citation metrics. Recently, approaches have shifted toward leveraging citation graph structures rather than relying solely on raw citation counts, for example with employing the PageRank method for measuring scientific prestige (Chen et al., 2023). Beyond ranking influence, PageRank also serves as a valuable tool for comparing bibliometric databases, revealing that citation-based prestige inherently depends on the completeness and accuracy of the chosen dataset. The PageRank computation for WoS (2000–2021) and OpenAlex (1950–2020) (Figure 1) highlights differences in both temporal coverage and citation network structure. Notably, no PageRank is calculated for the most recent 10 years in either dataset, as the metric requires a 10-year citation window. Beyond this temporal aspect, the results also reveal structural variations between the two citation networks.

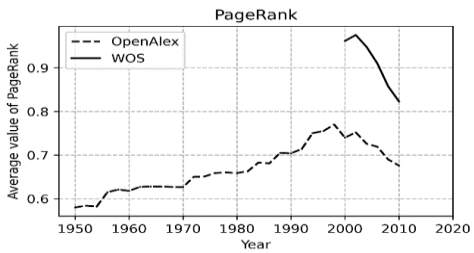


Figure 1. The structural differences of bibliometric datasets illustrated by a PageRank metric following Chen et al. (2023) with 10 years citation span and a damping factor of 0.5 on WoS and OpenAlex.

Useful Clustering Techniques for multi-disciplinary publication data

Bibliometric analysis of large datasets like WoS or OpenAlex requires automated classification of articles by topic. Determining the appropriate number of clusters is challenging, as disciplines do not always have clear boundaries, and articles often span multiple subjects, which necessitates a multi-label classification approach.

Our multi-labeling approach leverages the graph structure of publication data, where references link articles, and similarity is defined through a distance function computed from this network (cf. Nepusz et al., 2008).

Given $S = (s_{ki}) \in \mathbb{R}^{N \times N}$ with $s \in [0,1]$ describing the similarity between articles i and j and

$X = (x_{ki}) \in \mathbb{R}^{C \times N}$ with $x \in [0,1]$,

$\sum_{k=1}^{C} x_{ki} = 1 \; \forall \, i \in \{1, \ldots, N\}$

(see Table 1 for illustration)

assigning articles to clusters, we define

$$f(X) = \sum_{i=1}^{N} \sum_{j=1}^{N} (s_{ij} - \sum_{k=1}^{C} x_{ki} x_{kj})^2$$

to measure how well the similarity S is represented by the clustering X. From this we can compute a clustering by computing $argmin_X f(X)$, which is a continuous, non-convex optimization problem of very large size, as N > 107 and C depending on the number of clusters representing a meaningful number of (sub-)fields or topics, typically between 100 and 500, such as the 252 subfields defined in the OpenAlex database.

Table 1. Structure of the multi-label cluster matrix X, where x_{ki} indicates the assignment of Article i to Cluster k.

X	Article 1	Article 2	…	Article N
Cluster 1	x_{11}	x_{12}	...	x_{1N}
Cluster 2	x_{21}	x_{22}	...	x_{2N}
…	. . .	. . .	. . .	. . .
Cluster C	x_{C1}	x_{C2}	...	x_{CN}

A GPU-based gradient descent method is employed to efficiently handle large citation graphs. A subgraph with 700,000 nodes and the full OpenAlex graph, consisting of 60 million nodes and over a billion edges, were prepared for analysis. To enhance efficiency, the method leverages the sparse structure of the connection matrix and parallelizes gradient descent using CUDA. This parallelization allows for simultaneous processing of graph segments, significantly reducing computation time. Initial tests show that the CUDA implementation clusters the 700k subgraph in just 30 seconds, demonstrating highly promising performance.

Conclusions and Outlook

This study analyzed the development of academic publishing in WoS and OpenAlex using PageRank and introduced an efficient

multi-label clustering method to assess the similarity of academic publications. The comparison of PageRank-based rankings in both databases highlighted structural differences in citation networks, emphasizing the impact of data coverage and indexing practices. To address the challenge of disciplinary overlap in publication classification, a GPU-accelerated multi-label clustering approach was developed, leveraging the graph structure of citation networks.

While academic publication databases like WoS and OpenAlex provide the best available models of scholarly communication, they do not fully capture the broader landscape of academic publishing, often exhibiting biases such as an overrepresentation of English-language research, the exclusion of certain publication formats (monographs, book chapters etc.), and a disciplinary bias tilted towards STEM fields. Recognizing these limitations, our analysis is built on these databases, with the understanding that inherent biases must be considered when interpreting the results.

Acknowledgments

This work has been co-funded by the European Union (European Regional Development Fund EFRE, fund number: STIIV-001) Certain data included herein are derived from Clarivate™ (Web of Science™). © Clarivate 2025. All rights reserved. We acknowledge the use of WoS through the Kompetenznetzwerk Bibliometrie. Supported via the German Competence Network for Bibliometrics funded by the Federal Ministry of Education and Research (Grant: 16WIK2101A).

References

Chen, Y., Koch, T., Zakiyeva, N., Liu, K., Xu, Z., Chen, C-h., Nakano, J. & Honda, K. (2023). Article's scientific prestige: Measuring the impact of individual articles in the web of science. *Journal of Informetrics*, 17, 101379.

Nepusz, T., Petróczi, A., Négyessy, L. & Bazsó, F. Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. E 77, 016107, 2008.

De Solla Price, D. J. (1963). *Little Science, Big Science*. New York Chichester, West Sussex: Columbia University Press.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv*. https://arxiv.org/abs/2205.01833

https://doi.org/10.51408/issi2025_170

Fundamental Foundations of Scientific Heritage Formation: The Evolution of Scientific Knowledge and the Possibility of Applying Scientometric Tools

Victor A. Blaginin[1], Elizaveta V. Sokolova[2]

[1]*v.a.blaginin@usue.ru*, [2]*sokolova_ev@usue.ru*
Ural State University of Economics, 8 Marta St./Narodnaya Volya 62/45, Yekaterinburg (Russia)

Introduction

The rapid growth in the volume of scientific publications and the increasing complexity of the scientific knowledge structure leads to the formation of a confusing, unstructured system which makes it difficult to identify and evaluate key academic theories. In this regard, it is relevant to develop methodological approaches which facilitate the identification of emerging scientific theories, analyze their evolution, and predict their potential to become the scientific heritage. In this study, a scientific theory is understood as any scientific knowledge that can theoretically become the scientific heritage of an author, scientific school, organization, country, etc.

Understanding the regularities of the formation and consolidation of scientific theories has not only theoretical but also applied significance, as scientific heritage determines the basis for further research and influences the strategic development of science. This issue is especially relevant in the context of national scientific priorities, particularly for Russia during a period of external constraints, when the formation and support of its own scientific traditions become critical for the sustainable development of scientific and technological sovereignty.

Materials and Methods

In order to understand the mechanism of scientific theory development, the key concepts of the philosophy of scientific knowledge evolution were analyzed. The methodological basis of this work is grounded in Karl Popper's concepts of falsificationism and the evolutionary approach (Popper & Keuth, 1935; Popper, 1979), Thomas Kuhn's paradigm shifts (Kuhn, 1997), Imre Lakatos' research programs (Lakatos, 1976), Paul Feyerabend's epistemological anarchism (Feyerabend, 2020), and Larry Laudan's research traditions (Laudan, 1978), which are widely recognized as fundamental theories. The focus was on the criterion of scientific progress, the nature of theory change, and the sustainability of scientific concepts. The analysis was conducted using the methods of philosophical reconstruction, comparative analysis, and graphical modeling of the dynamics of scientific knowledge.

Results

Scientific Theory Development and the Network Path of Its Evolution

This study developed an original theoretical scheme of scientific theory development, including those studied earlier. According to the proposed model, a new scientific theory, even if initially formed in the works of one author, enters the scientific discourse rapidly, reaching a critical mass of recognition that corresponds to the phase of scientific revolution (T. Kuhn). From this point on, the theory develops and can give rise to offshoots – subtheories. At the same time, the original theory continues to exist and develop independently of these offshoots. Its further development follows the trajectory of successive falsifications and modifications (K. Popper). As critical anomalies accumulate, the theory reaches a bifurcation point, where it either faces the final crisis of the paradigm or a “protective belt” of auxiliary hypotheses is formed (I. Lakatos), supporting the main theory until the moment of anomalies re-accumulation, when variations are repeated cyclically (Figure 1).

Regarding the network path of scientific theory development, in the positive scenario,

the theory diversifies into many offshoots during its development, which nevertheless retain the rigid core of the original concept. At this stage, the factor of the scientific community manifests itself: competing interpretations and modifications (Feyerabend) contribute to the formation of a related sub-theories network, the stability and evolution of which are determined by the mechanisms of natural selection of ideas (late K. Popper).

The future fate of the entire theory depends on a number of factors: the depth of its embeddedness in the academic environment, the degree of diversification, the time interval of development, and resistance to external challenges.

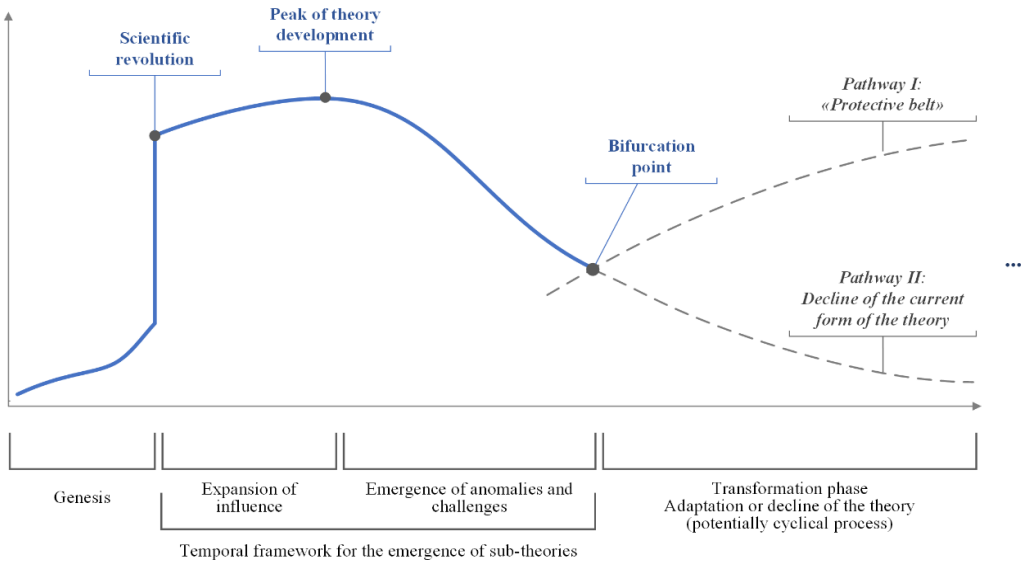


Figure 1. Conceptual framework of the development of an initial scientific theory.

At the same time, the model allows for the influence of exogenous factors – technological, social, and political transformations that can radically change the trajectory of scientific progress.

Methodology for Assessing the Dynamics and Determining the Status of Theories

The authors consider an approach to assessing scientific theories for their potential transition to the status of scientific heritage by analyzing the structure of the theory development network, in which the key indicator is the volume of citations – both for the theory itself and for the works that refer to it.

The authors suggest that a publication should be considered part of a sub-theory if it refers not only to works from the central (or previous) sub-theory but also to other works within the network. Thus, the analysis of the developmental network should include both top-down links from the core and horizontal links between nodes at different levels, which correspond to the complexity of the evolution of scientific knowledge.

Further analysis includes identifying regularities in the dynamics of the theory: the definition of its stages of evolution, critical points, and factors affecting its sustainability. In addition, based on the analysis of the current dynamics of the theory's development, including citation rates, researchers' activity, the emergence rate of new sub-theories, and the degree of their integration into scientific discourse, it is possible to assess the likelihood of its further successful development and potential transition to the category of scientific heritage.

Discussion

The authors intend to test the developed theoretical model using scientometric tools aimed at verifying the proposed scheme and assessing its applicability, which is reasonable in the context of the existing experience with scientometric analysis of network structures.

It is also envisaged that criteria will be developed for assessing the current stage of scientific theory evolution, based on the identification of general regularities and structural patterns inherent in the theories under study.

The model is to be validated through the application of bibliometric mapping, cohort analysis, citation path analysis, and network analysis, aimed at identifying the evolutionary phases of a theory, the structural interconnectedness of its sub-theories, and the degree of institutionalization of its conceptual core.

References

Popper, K.R., & Keuth, H. (1935). Logik der Forschung. *The Journal of Philosophy*, 32, 107.

Popper, K. R. (1979). *Objective knowledge: An evolutionary approach* (Vol. 49). Oxford: Clarendon press.

Kuhn, T. S. (1997). *The structure of scientific revolutions* (Vol. 962). Chicago: University of Chicago press.

Lakatos, I. (1976). Falsification and the Methodology of Scientific Research Programmes. *Can Theories Be Refuted?* 205–259.

Feyerabend, P. (2020). *Against method: Outline of an anarchistic theory of knowledge*. Verso Books.

Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth* (Vol. 282). Univ of California Press.

https://doi.org/10.51408/issi2025_210

Fusing Multi-Source Data through a Multi-Layer Network for Technological Opportunity Identification

Jinzhu Zhang[1], Mingxia Lu[2], Haoyu Li[3]

[1]*zhangjinzhu@njust.edu.cn*, [2] *lmxluna@163.com*, [3] *lhaoyu@njust.edu.cn*
Nanjing University of Science and Technology, Department of Information Management, Xiaolinwei Street 200, Nanjing (China)

Introduction

Predicting potential technology opportunities from vast data has been an indispensable research topic (Wang et al., 2023). With data volume and sources growing, identifying these opportunities has become much harder. Therefore, efficient methods for identifying technological opportunities are needed to recognize them accurately and comprehensively.

Previous research has shown that academic papers reveal foundational research topics (Jiang, Yang, & Gao, 2024), patents indicate emerging technology opportunities (Ba et al., 2024), and reviews reflect market-driven innovation potentials (Choi & Kwon, 2023). All of them are used in the research of technological opportunity identification. However, there are still problems with using multi-source data. Some scholars simply splice the texts of multi-source data, construct a single-layer network to extract combinations of knowledge units as technological opportunities. This approach fails to give sufficient consideration to the unique contributions of different data sources. While other scholars construct networks for each data source separately to identify technological opportunities, and then select the common knowledge unit pairs as the finally identified technological opportunities. This approach doesn't fully take into account the integration relationships among different data.

The emergence of multi-layer network theory offers a promising solution to this challenge. A multi-layer network can simultaneously display intra-layer and inter-layer relationships. By regarding each data source as a layer of the network and then constructing a three-layer network to combine the data from the three sources, we can easily integrate multi-source data with the help of the multi-layer network. This structure naturally links basic research, patented technologies, and market feedback, enabling each layer to contribute its unique information while facilitating cross-domain knowledge integration, thereby helping to identify more comprehensive technology opportunities.

This paper proposes a multi-layer network to integrate multi-source data for technological opportunity identification. Specifically, we extract a unified key phrase set from multi-source data, followed by constructing a multi-layer network based on the distinct co-occurrence patterns of these phrases within each data source. The unified phrase set ensures holistic utilization of multi-source information, while the multi-layer network preserves the inherent characteristics of individual data types. By integrating intra-layer relationships (reflecting domain-specific knowledge) and inter-layer connections (bridging cross-domain interactions), this method achieves synergistic integration of multi-source data and explicitly captures their technical interdependencies, thereby enabling comprehensive identification of technological opportunities.

Data and method

We conduct experiments using data from patents, academic papers, and consumer reviews in the field of new energy vehicles. Firstly, key phrases representing technological elements are extracted from each data source using tailored methods, while semantic unification is performed to address differences in expression across data types, constructing a unified multi-source key phrase set. Secondly, a multi-layer network is built by analyzing the co-occurrence relationships of the multi-source key phrase set across

different network layers, followed by network analysis. Finally, GCN and link prediction are employed to identify technological opportunities within the constructed multi-layer network.

Data description

This study selects data from the new energy vehicle sector for the year 2023 as the research subject. Patent data, scientific papers, and user reviews, including titles and abstracts (with review text referring to the content of the reviews), were collected from the Derwent Innovations Index, the SCIE database in Web of Science, and the Edmunds website (Edmunds.com), respectively. The search query was set as TS = ("new energy vehicle*" or "NEV$"), covering the period from Jan 1, 2023, to Dec 31, 2023. A total of 2,437 patent records, 758 academic papers, and 1,790 consumer reviews were retrieved.

Extraction and Fusion of Multi-Source Technical Knowledge Units

For patents and papers, TF-IDF extracts high-frequency keywords from texts as knowledge elements. RAKE and KeyBERT respectively capture syntactic and semantic features, with merged results yielding technical elements combining frequency, syntax, and semantics. For product reviews, BERTopic performs topic modeling to extract topical keywords. Syntactically and semantically salient phrases from RAKE/KeyBERT are deduplicated and integrated, deriving topic-relevant core phrases with syntactic-semantic features.

Finally, cosine similarity is used to unify synonymous key phrases. A higher similarity threshold (0.8) is set for patent and paper key phrases, while a lower threshold (0.5) is set for user review key phrases to capture more diverse and colloquial expressions.

Construction of Multi-Layer Networks

Technical elements in multi-source data exhibit diverse relationships. A three-layer network is constructed based on the co-occurrence relationships of key phrase sets in the three data types. This network is a multiplex network. The nodes in each layer are the same, all being multi-source key phrase sets, but the edges are different, representing different co-occurrence relationships in each data source.

Next, the edge overlap ratio metric is established to analyze the multi-layer network. The edge overlap ratio refers to the probability of overlapping connections between two network layers, indicating the inter-layer correlation between them. A higher overlap ratio suggests greater similarity in network structure, reflecting stronger inter-layer relationships. This metric measures the proportion of overlapping edges shared by two network layers, thereby reflecting their inter-layer structural similarity.

Identification of Technological Opportunities in Multi-Layer Networks

First, node embeddings are generated by integrating node features and local structural information through GCN, capturing semantic features and technical relationships of key phrases. Second, cross-source interaction is modeled by calculating association strength between data sources via edge overlap ratio, with weighted embeddings from support layers. Finally, link prediction is performed using optimized node embeddings to evaluate potential technical opportunities between key phrases.

In this study, positive samples are real key phrase pairs from the dataset, while negative samples are generated through random sampling. The dataset is split into training, validation, and test sets at an 8:1:1 ratio. Using the patent layer as the target domain and integrating papers and reviews as support layers, the model outputs link prediction scores reflecting technical association strength.

Result

We employed the accuracy, F1 score and AUC as evaluation metrics to assess the effectiveness of link prediction using only patent data and using multi-source data in a multi-layer network. The results are shown in Table 1, indicating that link prediction using the multi-source data multi-layer network has certain improvements in various indicators. The increase in the AUC value indicates that this method is better in the ability to distinguish between positive and negative samples; the rise in accuracy reflects the enhanced accuracy and reliability of the

prediction results; the improvement of the F1 value further proves the enhancement of the comprehensive performance of this method.

Table 1. The effectiveness of different methods.

	Accuracy	F1	AUC
Patent Data	56.67%	69.77 %	95.36 %
MSML	62.04%	72.46 %	96.00 %

Conclusion

This paper enriches the data sources for technological opportunity identification and applies the methods of deep learning and complex networks to make full use of multi-source information through a multi-layer network approach. Compared with single-data-source methods, this method performs better. In the next step, we plan to apply other deep-learning methods, which may perform more outstandingly in the semantic representation of multi-source phrases.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 72374103, 71974095), China Society of Indexers (Grant No. CSI24C10), and Jiangsu Provincial Federation of Philosophy and Social Sciences (Grant No. 24SYC-023).

References

Wang, J., Zhang, Z., Feng, L., Lin, K.-Y., & Liu, P. (2023). Development of technology opportunity analysis based on technology landscape by extending technology elements with BERT and TRIZ. *Technological Forecasting and Social Change*, 191, 122481.

Jiang, M., Yang, S., & Gao, Q. (2024). Multidimensional indicators to identify emerging technologies: Perspective of technological knowledge flow. *Journal of Informetrics*, 18(1), 101483.

Ba, Z., Meng, K., Ma, Y., & Xia, Y. (2024). Discovering technological opportunities by identifying dynamic structure-coupling patterns and lead-lag distance between science and technology. *Technological Forecasting and Social Change*, 200, 123147.

Choi, K. H., & Kwon, G. H. (2023). Strategies for sensing innovation opportunities in smart grids: In the perspective of interactive relationships between science, technology, and business. *Technological Forecasting and Social Change*, 187, 122210.

https://doi.org/10.51408/issi2025_168

Gender Disparities in Editorial Board Member in Information Science & Library Science Journals

Yiming Liu[1], Rut Lucas-Domínguez[2], Adolfo Alonso-Arroyo[3], Cristina Rius[4], Rafael Aleixandre-Benavent[5]

[1]*Yiming.Liu@uv.es*, [3]*adolfo.alonso@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)

[2]*rut.lucas@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)
CIBERONC (Spain)

[4]*crisina.rius@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)
National Centre for Cardiovascular Research (CNIC) (Spain)
CIBERCV (Spain)

[5]*rafael.aleixandre@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Ingenio (CSIC-Polytechnic University of Valencia) (Spain)

Introduction

Members of editorial boards of scientific journals (EBMs) are qualified professionals characterized by their professional achievements and academic leadership (Kaji et al., 2019). Gender studies of EBMs are of great interest because of their impact on both equity and the quality and representativeness of science (Mauleón et al., 2013). In medicine, for example, the proportion of women in EBMs increased from 11% to 23% between 1993 and 2003 (Keiser et al., 2003). In Mathematics, it was 8.9% (Topaz & Sen, 2016), and in some Information Science & Library Science (ISLS) journals, it was 40% (Willett, 2013). Thus, it is evident that there is a gender disparity in EBMs in some fields, with lower representation of women, especially as Editors-in-Chief (EiC). The current representation of women as EBMs in ISLS journals is unknown, as well as their geographical origin and their roles within the EBMs. Therefore, the objectives of this work are: a) to determine whether there are gender differences in the EBMs of ISLS journals according to their country of origin, quartile in Journal Citation Reports (JCR), and role in the committees; b) to analyze the representation of women in the journal editorial boards.

Methodology

The EBMs of 83 journals in the ISLS category of the JCR, Social Science Citation Index edition of 2023, were identified. To classify the different roles of the EBMs, the methodology of Liu et al. (2023) was

followed, which grouped them into 10 categories. Each journal was assigned the publisher and the quartile in JCR (Q), and each EBM was assigned the gender (female or male) and its country of affiliation. Figure 1 shows the flowchart of the methodological process.

Results

A total of 4,468 EBMs from 86 different countries were identified, 2,813 males (62.96%) and 1,616 females (36.17%). The gender of 39 EBMs (0.87%) could not be determined. The United States has the largest number of EBMs (n=1,697), followed by the United Kingdom (n=445). The percentage of women
in the 10 countries with the highest number of EBMs is less than 45% in all of them, except in South Africa (58.33%), Malaysia (57.69%) and Israel (52%). Female representation is lower in China (25.36%), Germany (29.55%), France (29.67%), Italy (32.94%), and Spain (33.33%) (Figure 2).

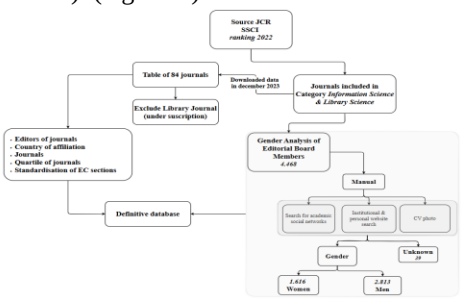


Figure 1. Flowchart for gender analysis of EBMs.

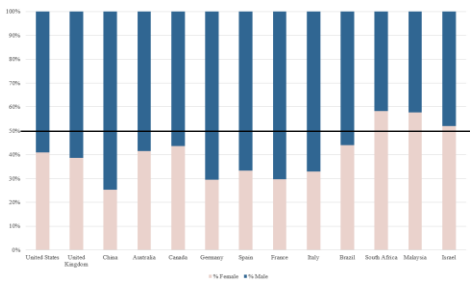


Figure 2. Gender distribution of EBMs by geographical origin.

70.07% of male EBMs (n=1,491) were found in Q1 of the journals, while the distribution of females follows a reverse staircase pattern from Q4 to Q1: $Q4_{Female}$=49.58%; $Q3_{Female}$=46.6%; $Q2_{Female}$=36.88%; and $Q1_{Female}$=29.93%. The representation of men is above 65% in all roles classified in Q1, except in the role "Administration, Marketing and Communication" (AMC), where it is 45%, and in addition, 79.17% of the EiC in Q1 are men. On the other hand, the percentage of women increased in Q4: 83.33% in the role of reviewer; 59.26% in the role of associate editor; 54.67% in the role of AMC; 50.91% in the role of editorial board; and 48.15% in the role of EiC (Figure 3).

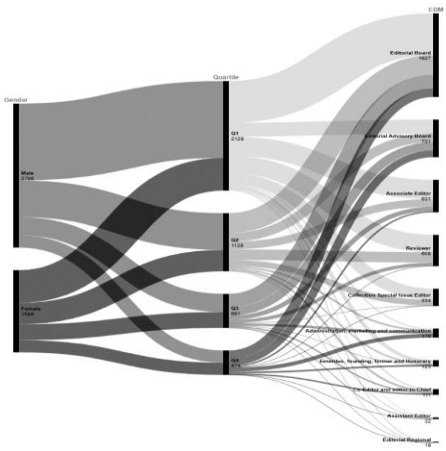


Figure 3. Sankey diagram of gender, journal quartile and role relationships in EBMs.

Among the 10 publishers producing more than 2 journals (Figure 4), Elsevier (n=11), Taylor & Francis (n=10), and Emerald Group Publishing (n=10) lead. Commercial publishers such as Igi Global (81.71%), Springer (70.78%) and Elsevier (69.63%) have a high proportion of men, while academic publishers such as Johns Hopkins University Press (65.38%), American Library Association (60.53%) and Oxford University

ress (43.41%) have a high proportion of women.

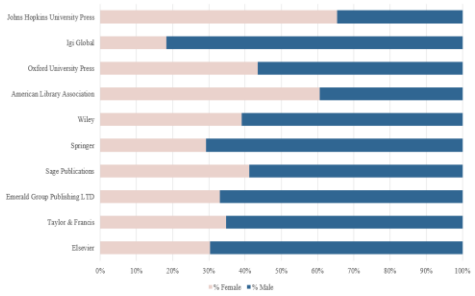


Figure 4. Gender distribution according to publishers publishing more than one journal.

Conclusion

There is a predominance of men among the EBMs of ISLS journals in most countries, and also among journals in the upper quartiles. The representation of women is inversely related to the quartile, increasing as the quartile decreases, and there are more female EiC in the lower quartiles than in the upper quartiles. Women are more represented in some specific roles, such as AMC. Only in two of the ten publishers with more than two journals does the percentage of women exceed 50%. This work confirms the existence of gender differences in the EBM of journals in the ISLS field, both in management positions and in most roles within each editorial board. As future work, it would be interesting to analyze the evolution of EBMs in the coming years, as well as to investigate the causes of the lower representation of women, including the analysis of historical and socio-cultural factors, power dynamics in the ISLS field, and social expectations that may affect the participation and leadership of women in EBMs.

Funding

Yiming Liu is a predoctoral research fellow under the Generalitat Valenciana Predoctoral Research Personnel Training Program (CIACIF/2023/316).

Acknowledgements

Betlem Ortiz Campos. Beneficiary of Call for Technical Support Staff Grants. Ministry of Science and Innovation. State Research Agency. Co-financed by European Union (PTA2021-019882-I).

References

Kaji, A.H., Meurer, W.J., Napper, T., Nigrovic, L.E., Mower, W.R., Schriger, D.L., et al. (2019). State of the Journal: Women first authors, peer reviewers, and Editorial Board Members at Annals of Emergency Medicine. *Annals of Emergency Medicine, 74*(6), 731-735. https://doi.org/10.1016/j.annemergmed.2019.05.011

Keiser, J., Utzinger, J., & Singer B.H. (2003). Gender composition of editorial boards of general medical journals. *The Lancet, 362*(9392), 1336. https://doi.org/10.1016/S0140-6736(03)14607-7

Liu, Y., Alonso-Arroyo, A., Aleixandre-Benavent, R. & Valderrama-Zurián, J.C. (2023). Editorial boards of Information Science and Library Science Journals: roles, terminology, origin, and internationalization. *Profesional de la Información, 32*(6), e320614. https://doi.org/10.3145/epi.2023.nov.14

Mauleón, E., Hillán, L., Moreno, L., Gómez, I. & Bordons, M. (2013). Assessing gender balance among journal authors and editorial board members. *Scientometrics, 95*(1), 87-114. http://dx.doi.org/10.1007/s11192-012-0824-4

Topaz, C.M. & Sen, S. (2016). Gender representation on journal editorial boards in the mathematical sciences. *Plos one, 11*(8), 1-21. https://doi.org/10.1371/journal.pone.0161357

Willett, P. (2013). The characteristics of Journal Editorial Boards in Library and Information Science. *International Journal of Knowledge Content Development & Technology, 3*(1), 5-17. https://doi.org/10.5865/IJKCT.2013.3.1.005

https://doi.org/10.51408/issi2025_195

Gender Leadership in Cancer Research

Cristina Rius[1], Yiming Liu[2], Adolfo Alonso-Arroyo[3], Rafael Aleixandre-Benavent[4], Rut Lucas-Domínguez[5]

[1]*crisina.rius@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)
National Centre for Cardiovascular Research (CNIC) (Spain)
CIBERCV (Spain)

[2]*Yiming.Liu@uv.es*, [3]*adolfo.alonso@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)

[4]*rafael.aleixandre@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Ingenio (CSIC-Polytechnic University of Valencia) (Spain)

[5]*rut.lucas@uv.es*
UISYS Research Group, Unit of Information and Social and Health Research, University of Valencia. Associated Unit to INAECU. Interuniversity Institute for Advanced Research on Science and University Evaluation. UC3M-UAM (Spain)
Department of History of Science and Documentation, University of Valencia (Spain)
CIBERONC (Spain)

Introduction

Cancer is a major global health problem. In 2022, there were an estimated 20 million new cancer cases and 9.7 million deaths, with projections indicating 35 million cancer cases by 2050. In Europe, cancer is the second leading cause of death, almost equal to cardiovascular diseases (WHO, 2024). Research progress in this area has traditionally been reported through scientific publications, the analysis of which has included bibliometric studies of the content and authorship of scientific activity, with a particular focus on the gender distribution of authorship. The aim of this study was to assess the progress made in incorporating a gender perspective in cancer research, comparing the years 2011 and 2021, using a dual approach that includes both the analysis of authorship in the publications derived and the evaluation of the scientific content of the research carried out according to the type of cancer studied.

Methods

Identification of cancer articles and retrieval of MeSH terms

A bibliographic search was carried out for articles and reviews in the field of oncology that were signed by at least one Spanish institution during the period 2011-2021 through the Science Citation Index Expanded database of the Web of Science Core Collection, which yielded 50,776 documents (Lucas-Domínguez et al., 2024). A PubMed/Medline search was then performed using the PMIDs of the retrieved records, which produced 47,940 papers. All MeSH

terms were then downloaded from the records and a total of 43,086 papers containing MeSH terms were identified (89.87% of the total number of papers indexed in PubMed/Medline). The retrieved records were exported to a relational database in Microsoft Access using in-house developed bibliometrics software.

Classification of papers by cancer type

Global cancer statistics indicate that the highest incidence is mainly due to lung, breast, colorectal, prostatic and stomach cancers. On the other hand, lung and colorectal cancer are the leading causes of death, followed by liver, breast and stomach cancer (Bray et al., 2024). The 43,086 papers containing MeSH terms were evaluated according to the above-mentioned cancer typologies using the specific representative descriptors obtained from the MeSH tree (Neoplasms by Site [C04.588]) (Table 1).

Table 1. MeSH descriptor analysis of retrieved cancer papers.

Cancer type	MeSH	Papers	%*
BREAST	Breast Neoplasms	3,417	93.6
	Triple Negative Breast Neoplasms	210	5.8
	Carcinoma, Ductal, Breast	192	5.3
COLORECTAL	Colorectal Neoplasms	1,969	62.7
	Colonic Neoplasms	626	19.9
	Rectal Neoplasms	409	13.0
LUNG	Lung Neoplasms	2,424	95.3
	Carcinoma, Non-Small-Cell Lung	1,193	46.9
	Adenocarcinoma of Lung	163	6.4
PROSTATIC	Prostatic Neoplasms	1,458	87.7
	Prostatic Neoplasms, Castration-Resistant	224	13.5
LIVER	Liver Neoplasms	1,276	96.4
	Carcinoma, Hepatocellular	816	61.7
	Liver Neoplasms, Experimental	35	2.6
STOMACH	Stomach Neoplasms	453	100.0

*The percentages do not add up to 100% as there may be more than one MeSH in a record

Gender analysis of authorship in cancer publications

To identify the signatures of the 43,086 papers, the authors' names were manually standardised and gender was assigned using the statistical package Genderize.io (https://genderize.io/#overview). The papers were then assigned to the following groups: gender parity (P), when the percentage of one of the genders was between 40% and 60% of the total number of authors signing the article; female majority (FM) and male majority (MM) authorship.

Results

The 43,086 retrieved articles on cancer that were signed by at least one Spanish institution were analysed using MeSH terms to classify them into the different types of cancer (Table 1). The description of the frequency of research on the different types of cancer in the articles is shown in Figure 1. As can be seen, publications on 6 cancers predominate: breast, colorectal, lung, prostatic, liver and stomach, demonstrating the correlation between the cancers with the highest incidence and mortality and the research carried out.

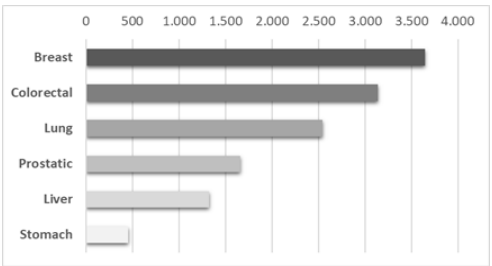


Figure 1. Analysis of the 6 most common cancers covered in oncology publications for the period 2011-2021.

The 12,272 articles corresponding to the 6 most studied cancer typologies are 24.5% of the cancer records retrieved for the entire period 2011-2021 (Table 2). Of these, 11,019 articles had all author signatures identified, highlighting the majority of male authorship in all the cancers studied, except for breast cancer, where the parity of signatories predominates. Comparing the years 2011 and 2021 and the participation in authorship by sex, 828 articles (123 FM, 253 P, 324 MM) and 1,388 articles (269 FM, 478 P, 504 MM) respectively were identified.

Table 2. Classification of papers by cancer type and gender of authors.

Cancer type	Papers	All author signatures identified	% Papers all author identified	Female Majority	% Female Majority	Parity	% Parity	Male Majority	% Male Majority
Breast	3,649	3,159	86.57%	847	26.81%	1,346	42.61%	966	30.58%
Colorectal	3,141	2,716	86.47%	566	20.84%	1,051	38.70%	1,099	40.46%
Lung	2,544	2,224	87.42%	272	12.23%	768	34.53%	1,184	53.24%
Prostatic	1,662	1,360	81.83%	195	14.34%	407	29.93%	758	55.74%
Liver	1,323	1,177	88.96%	187	15.89%	389	33.05%	601	51.06%
Stomach	453	383	84.55%	76	19.84%	144	37.60%	163	42.56%

The analysis of specific authorship groups by cancer type shows that in 2021, compared to 2011, there is a slight trend towards an

increase in papers with parity compared to a decrease in papers with male majority authorship, except for liver cancer (Figure 2). In contrast, the increase in female majority authorship is minimal, except for colorectal cancer and stomach cancer.

Conclusion

The integration of gender equality in science remains a critical issue despite various socio-political initiatives across Europe and global commitments, such as the 2030 Agenda and the United Nations System-wide Plan of Action for Gender Equality and the Empowerment of Women (UN-SWAP). These efforts, including those by the World Health Organization, emphasize gender mainstreaming in research. However, significant challenges persist, highlighting the need for continued and expanded actions to address the structural and cultural barriers that hinder women's full participation in science and decision-making in research and innovation. Urgent efforts are needed to achieve true gender equity and human rights integration in scientific policies and practices (Rius et al., 2024).

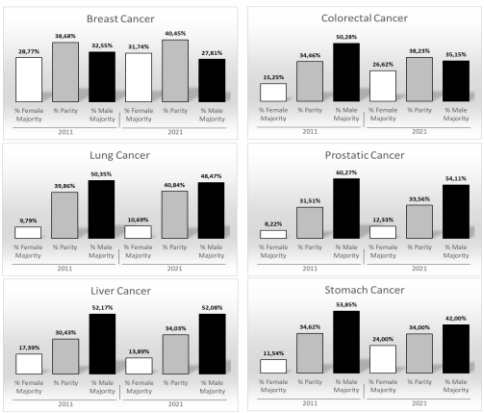


Figure 2. Gender gap in cancer research between 2011 and 2021.

Acknowledgments

Spanish Ministry of Equality (MUJER-PI-21-3-ID24). Valencian Regional Ministry of Innovation, Universities, Science, and Digital Society. Generalitat Valenciana (CIAICO/2021/205). Betlem Ortiz Campos. Technical Support Staff Grants. Ministry of Science and Innovation. State Research Agency. Co-financed by the UE (PTA2021-019882-I). Yiming Liu. Predoctoral Training Programme of the Generalitat Valenciana (CIACIF/2023/316).

References

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3), 229-263.

Lucas-Domínguez, R., Aragonés González, M., Sixto-Costoya, A., Ruiz-Martínez, E., Alonso-Arroyo, A., & Valderrama-Zurián, J. C. (2024). The inclusion of the gender perspective in oncology research with Spanish participation. *Heliyon*, 10(9), e30043.

Rius, C., Sixto-Costoya, A., Lucas-Domínguez, R., & Valderrama-Zurián, J. C. (2024). State-of-the-Art on Gender Equality in Cardiovascular Research. *Women's Health Reports*, 5(1), 897-908.

World Health Organization. (2024). Global cancer burden growing, amidst mounting need for services. Retrieved April 30, 2024 from: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services.

https://doi.org/10.51408/issi2025_174

Healthcare Cybersecurity: Insights from a Scientometric Approach

Simone Di Leo[1], Cinzia Daraio[2], Fabio Nonino[3], Eugenio Oropallo[4]

[1]*dileo@diag.uniroma1.it*, [2]*daraio@diag.uniroma1.it*, [3]*fabio.nonino@uniroma1.it*, [4]*eugenio.oropallo@uniroma1.it*

Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), Sapienza University of Rome, Via Ariosto 25, Rome, 00185 (Italy)

Introduction

The increasing digitization of healthcare, driven by technological advancements and the pursuit of enhanced patient care, presents both unprecedented opportunities and significant cybersecurity challenges. While digital tools, patient phygital twins for medical planning and connected devices streamline processes and improve access to care, they simultaneously expand the attack targets for malevolent actors, potentially compromising sensitive data and patient safety (Spanakis et al., 2020). Existing technical cybersecurity countermeasures aim to protect the confidentiality, integrity, and availability of healthcare data and information systems, but the rising frequency and sophistication of cyberattacks necessitate a deeper understanding of the evolving threat landscape (Jalali et al., 2019). The SEcurity and RIghts in the CyberSpace (SERICS) project is currently developing remote healthcare solutions based on personal devices while the Phygital Twin Technologies for Innovative Surgical Training & Planning project is developing a phygital twin device and software for surgical planning, further highlighting the critical need for robust cybersecurity measures. Connected medical devices and electronic health records (as done for the patient phygital twin), while offering substantial benefits, introduce new vulnerabilities that require careful consideration. Effective incident response strategies are crucial for healthcare organizations to mitigate the impact of cybersecurity incidents and ensure timely recovery. This study addresses the critical need for robust cyber defences and effective response processes within the healthcare sector, emphasizing their contribution to overall cyber resilience through adherence to industry best practices. This is a macro-level analysis of cyber incidents across different countries and cyber actors–which aims to identify frequently targeted entities and prominent threat actors within the healthcare ecosystem. This analytical approach, leveraging real-world incident data, provides a valuable contribution by uncovering systemic vulnerabilities and informing targeted cybersecurity strategies within the context of the SERICS project and the broader healthcare landscape.

Data

This study utilizes data from the–European Repository of Cyber Incidents (EuRepoC) database (https://eurepoc.eu/), providing a comprehensive dataset of cyber incidents from 2000 to present, with ongoing daily data collection and curation. Data specific to healthcare cyber incidents, retrieved from EuRepoC (Version 1.2), shows a total of 348 incidents and 129 actors involved.

Method

Employing the methods of scientometrics (Garfield, 1972, 1955; Marchiori, 1997), the HITS algorithm (Kleinberg, 1999) analyzes networks by assigning two scores to each node: *authority* (representing value as a source of information) and *hub* (representing value as a curator or aggregator of information). Formally, these scores are defined as follows:

- *Authority Update Rule*: $\text{auth}(p) = \sum \text{hub}(i)$, where i represents all nodes that link to node p. So, the authority score of a node p is the sum of the

hub scores of all nodes pointing to it.

- *Hub Update Rule*: hub(p) = $\sum$auth(i), where i represents all nodes that node p links to. So, the hub score of a node p is the sum of the authority scores of all nodes it points to.

The algorithm begins by initializing each node with both hub and authority scores of 1. It then iteratively updates these scores using the above formulas. After each iteration, the scores are normalized to prevent unbounded growth. The algorithm differentiates between individual interactions of two actors by representing each as a weighted arc. This differentiation subsequently affects the derived authority and hub scores. When applied to a network of healthcare cyber incidents, the authority score reflects how often an actor is targeted (a "defender" score), while the hub score reflects how often they initiate attacks (an "aggressor" score).

Results and conclusion

The analysis of cyber incidents within the healthcare sector yielded significant insights into the landscape of cyber threats. For the sake of brevity, we only report the results of the top 10 authorities and hubs actors. Authority results (Table 1) highlighted the United States (0.5424), Japan (0.3471), and Israel (0.3398) as prominent targets within the healthcare sector. Conversely, Hub scores (Table 2) revealed the Democratic People's Republic of Korea (0.8046) with a notably high score, followed by Iran (0.3625) and China (0.3596). Beyond nation-state actors, the analysis identified criminal groups such as CosmicBeetle, TA558, and various ransomware groups (Rhysida, LockBit, BianLian, BlackCat/ALPHV) as significant hubs, underscoring the complex and multifaceted nature of the cyberattack landscape affecting healthcare, as initially emphasized by the increasing digitization and its associated risks. These findings directly contribute to the aims of the SERICS project. By identifying frequently targeted entities and prominent threat actors, this research provides crucial information for the development of robust cybersecurity measures within the previously mentioned projects. The identification of specific threat actors and their tactics informs the design and implementation of targeted security protocols for personal, and devices used in remote healthcare, mitigating the vulnerabilities introduced by connected medical devices and sensible electronic health records. This preliminary analysis underscores the need for further research to explore correlations between Authority and Hub scores, analyze the temporal evolution of these metrics. The analysis reveals key insights with significant implications for policy development. Current cybersecurity frameworks often prioritize organizational-level security measures. However, this analysis suggests that effective policy must operate at multiple levels simultaneously, recognizing the crucial role of international cooperation. Drawing on the concept of "networked governance," a policy approach that acknowledges and addressesthe interconnected nature of cyber threats, as proposed by Eggers and Goldsmith (2004), is essential.

Table 1. Top 10 Authorities.

Actor	Authority score
United States	0.5424
Japan	0.3471
Israel	0.3398
Korea, Republic of	0.3258
United Kingdom	0.2956
Germany	0.2551
Spain	0.2046
France	0.2027
Russia	0.1845
China	0.1838

Table 2. Top 10 Hubs.

Actor	Hub Score
Korea, Democratic People's Republic of	0.8046
Iran, Islamic Republic of	0.3625
China	0.3596
Russia	0.1051
CosmicBeetle	0.0902
TA558	0.0866
Rhysida Ransomware Group	0.076
LockBit	0.0757
BianLian Ransomware Group	0.0727
BlackCat/ALPHV	0.0727

Acknowledgments

The poster is part of the Research Projects "Phygital Twin Technologies for Innovative Surgical Training & Planning (ECS 0000024 – Rome Technopole)" and is partially supported by project SERICS (ECS B53C22003990006) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU.

References

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. Science, 178(4060), 471-479.,

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. Science, 122(3159), 108-111.

Marchiori, M. (1997). The quest for correct information on the Web: Hyper search engines. Computer Networks and ISDN Systems, 29(8-13), 1225-1235.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. ACM computing surveys (CSUR), 31(4es), 5-es.

Spanakis, E. G., Bonomi, S., Sfakianakis, S., Santucci, G., Lenti, S., Sorella, M., ... & Magalini, S. (2020, July). Cyber-attacks and threats for healthcare–a multi-layer thread analysis. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 5705-5708). IEEE.

Jalali, M. S., Russell, B., Razak, S., & Gordon, W. J. (2019). EARS to cyber incidents in health care. Journal of the American Medical Informatics Association, 26(1), 81-90.

Eggers, W. D., & Goldsmith, S. (2004). Government by network. The New Public Management Imperative.

https://doi.org/10.51408/issi2025_209

How Does Knowledge Source Novelty Influence Knowledge Output Novelty? Evidence from 269,569 PLOS Articles

Yi Xiang[1], Chengzhi Zhang[2]

[1] xiangyi@njust.edu.cn, [2] zhangcz@njust.edu.cn
Department of Information Management, Nanjing University of Science and Technology, 210094, Nanjing (China)

Introduction

Novelty is a key criterion in evaluating the innovativeness of academic research. As academic literature expands rapidly, effectively measuring novelty has become a critical research focus. Existing methods for assessing the novelty of academic papers can be classified into two categories based on the knowledge components they employ: (1) citation-based methods and (2) word-level knowledge unit methods, such as MeSH terms and entities. Citation-based approaches capture novelty in knowledge sources, while entity-based approaches focus on research content. However, prior studies often examine these methods in isolation, neglecting their interconnections. Investigating their relationship can deepen our understanding of measurement discrepancies and correlations, providing a theoretical basis for integrating multiple novelty dimensions to improve accuracy.

Addressing the limitations of existing research, this study models academic paper writing as a production process. We then apply the Cobb-Douglas production function (Douglas, 1928), commonly used in economics to model the relationship between input and output, to examine the relationship between the novelty of knowledge sources and the novelty of knowledge output in academic papers.

Methodology

Dataset

We collected 362,269 papers published between 2003 and September 2024 from the PLOS database. After extracting reference records and analysing corresponding journals, we excluded papers with missing reference lists, resulting in a final dataset of 330,966 papers. We then retrieved MeSH term lists from OpenAlex[1] and excluded records with missing data, yielding 269,569 papers. As MeSH terms pertain to biomedical fields, this filtering indicates that the study focuses primarily on biomedical literature. A basic statistical analysis of the dataset revealed that, on average, each paper cites 23 different journals and contains 17 MeSH terms.

Novelty Measurement

We propose a graph representation learning approach to measure novelty, based on combinatorial innovation theory (Uzzi et al., 2013). For papers published in year Y, we first compile prior papers, extracting knowledge components (reference journals or MeSH terms) as network nodes, with edges linking co-occurring units. We then apply the LINE algorithm (Tang et al., 2015) to generate vector representation of nodes.

Given a focal paper with N knowledge units, each represented by a vector V_i, we construct all possible knowledge unit combinations. The novelty of each combination $Comb_{i,j}$ is then quantified using the following formula:

$$Novelty_{i,j}^{comb} = 1 - \frac{|V_i||V_j|}{V_i \cdot V_j} \#(1)$$

The overall novelty of the paper is the sum of the novelty scores for all combinations. Since this study considers two types of knowledge units—references and MeSH terms—we distinguish between them by denoting reference-based novelty as $Novel_J$ and MeSH-based novelty as $Novel_M$.

[1] https://openalex.org/

Cobb-Douglas production function

The Cobb-Douglas function is widely used in economics to model the relationship between inputs (e.g., capital and labor) and outputs in production activities. The writing of academic papers can also be viewed as a production process, where scholars accumulate raw experience by reading references, invest time and effort to validate research ideas, and ultimately produce research papers. Therefore, we consider the novelty of references $Novel_J$ as capital input, and the number of authors (L) as labor input. The novelty based on MeSH terms $Novel_M$ serves as the output. We then model the relationship between these variables using the transcendental logarithmic model (Christensen et al., 1973), an extension of the Cobb-Douglas function that accounts for interactions between input factors:

$$\begin{aligned} ln\ ln\ Novel_M = & ln\ ln\ A + \alpha\ ln\ ln\ Novel_J \\ & + \beta\ ln\ ln\ L \\ & + \gamma\left(ln\ ln\ Novel_J \cdot ln\ ln\ L\right) + \varepsilon \#(2) \end{aligned}$$

Where, A and ε represent the intercept and the error term, respectively.

Result

Descriptive Statistics

Figure 1 illustrates the distribution of paper novelty calculated using two methods: MeSH term-based and reference-based novelty. Both methods reveal a clear right-skewed distribution, indicating that the majority of papers exhibit low novelty, while only a small proportion are classified as highly novel.

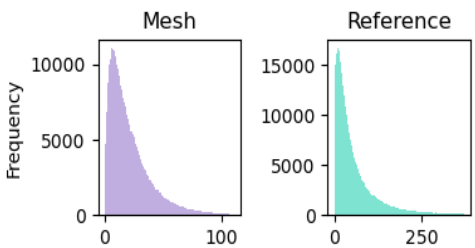


Figure 1. The distribution of papers' novelty.

Analysis of the relationship between two types of novelty

Table 1 presents the regression results based on Equation (2). First, regarding the novelty of knowledge sources, when the number of authors is held constant, each unit increase in $Novel_J$ is associated with an average increase of 0.0359 in the novelty of the output $Novel_M$. The impact of the number of authors (L) on output novelty is even more pronounced. For each additional author, $Novel_M$ is expected to increase by 0.5133, consistent with prior research. We also examined the quadratic term for the number of authors and found its coefficient to be negative, suggesting that beyond a certain threshold, additional authors may diminish output novelty. Furthermore, the interaction term between $Novel_J$ and L has a negative effect on the dependent variable, indicating that the influence of knowledge source novelty and the number of authors on output novelty may counteract each other.

These findings demonstrate that the novelty of knowledge sources positively influences the novelty of a paper's content. However, the number of authors also plays a crucial role in knowledge flow. The negative interaction between $Novel_J$ and L suggests that while an increase in the number of authors may introduce diverse perspectives and enhance novelty, excessive collaboration can lead to higher coordination costs. Additionally, researchers from different backgrounds may have varying perceptions of novelty, potentially hindering the effective translation of knowledge source novelty into novel research output.

Table 1. The regression results.

	(1)	(2)
$Novel_J$	0.0359***	0.0279***
	(0.004)	(0.004)
L	0.5133***	0.7944***
	(0.008)	(0.014)
L^2		-0.0871***
		(0.004)
$Novel_J * L$	-0.0275***	-0.0241***
	(0.002)	(0.002)
Constant	1.8048***	1.6166***
	(0.015)	(0.017)
Observations	269,569	269,569
Pseudo R^2	0.061	0.063

Note: ***: p<0.001.

Conclusion

This study examines the relationship between the novelty of knowledge sources (references) and outputs (MeSH terms) in academic papers. We propose a graph representation learning method to measure novelty and use the Cobb-Douglas function to model idea transformation as a production process. Findings reveal that source novelty significantly impacts output novelty, advancing our understanding of knowledge flow. However, factors such as team diversity and funding may influence this relationship. Future research should explore these variables and assess the findings' generalizability across disciplines.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.72074113).

References

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1973). Transcendental logarithmic production frontiers. The review of economics and statistics, 28-45.

Douglas, P. (1928). Cobb douglas production function. The Quarterly Journal of Economics, 42(3), 393-415.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. Science, 342(6157), 468-472.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015, May). Line: Large-scale information network embedding. In Proceedings of the 24th international conference on world wide web (pp. 1067-1077).

https://doi.org/10.51408/issi2025_185

How far are we from understanding phygital healthcare convergence? Building an AI knowledge map grounded in bibliometric metadata

Cinzia Daraio[1], Simone Di Leo[2]

[1]*daraio@diag.uniroma1.it*, [2]*dileo@diag.uniroma1.it*
Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), *Sapienza* University of Rome, Via Ariosto 25, Rome, 00185 (Italy)

Introduction

Initially prevalent within the marketing domain, the increasing convergence of the physical and digital realms, commonly termed the "*phygital*" concept (Batat, 2024), is revolutionizing various sectors, including healthcare. This influence is exemplified by prominent initiatives in this field such as the Rome Technopole project "*Phygital Twin Technologies for Innovative Surgical Training & Planning"* using the *phygital* concept by developing digital and physical technologies for planning and training surgeons. The *phygital* paradigm creates hybrid experiences blending real-world interactions with virtual ones, offering new opportunities for learning and practice. This can be done using various technologies like Digital Twins (DTs), Augmented Reality (AR), Virtual Reality (VR), and Artificial Intelligence (AI). DTs and virtual representations of physical entities, processes, or systems, are key elements in this context, providing simulated environments for training, planning, and analysis. However, the *phygital* concept extends beyond DTs, encompassing a broader ecosystem of technologies. This research delves into this ecosystem using an innovative methodology that combines traditional bibliometric metadata (author's keywords) with the power of AI to create a knowledge map (KM) on the *phygital* concept in healthcare to find convergence relationships of the different concepts. KMs are graphical representations of knowledge, where nodes represent concepts and edges represent relationships between them. KMs can help researchers to identify areas and research landscapes within a specific field or topic. Bibliometric analysis, which has traditionally relied on citation counts and co-occurrence analysis, is commonly used to understand research landscapes. However, this approach has various limitations (see e.g. Haustein & Larivière, 2014). This research addresses this gap by utilizing AI-generated KM while concurrently preserving bibliometric source metadata. To create the KM, we leverage the capabilities of the AI by Google called Gemini 2.0 (G2, https://gemini.google.com/app). Base the KM creation to the bibliometric metadata ensures the reliability of the generated knowledge by grounding it in verifiable data sources and mitigating potential biases that may arise from exclusive reliance on AI-
only information. This approach offers a more comprehensive perspective than conventional bibliometric methods by enabling the identification of prominent research areas, subtle relationships and emerging subfields on the *phygital* in the healthcare context.

Method

This study employed a two-stage process to construct the KM of the *phygital* concept within the healthcare field. The first stage involved a systematic search within the Scopus database (https://www.scopus.com/) to retrieve the bibliometric metadata. The query was conducted on the TITLE-ABS-KEY fields. The search combined keywords related to digital representations ("Digital twin", "Digital twins", "Digital phantom", "Digital phantoms", "phygital" and "physical twin"), training applications ("practice", "training", "teaching", "didactic", "Didactic purpose", "Surgical training" and "Robot assisted surgery"), and the healthcare domain

("healthcare", "health care" and "medical"). The search was done on January 16, 2025. From the retrieved articles, author-supplied keywords were extracted. The second stage used the retained keywords as input for G2, using *Termboard* (https://termboard.com/app#/ generated prompt) to identify the relations between these keywords and the *phygital* concept. To contextualize the analysis, the definition of *phygital* provided in the Introduction Section was explicitly provided to G2 before the identification of the relations. This step was crucial to guide the AI's exploration of the connections within the extracted keyword set, assisting the construction of the KM.

Results and conclusion

From the Scopus search, we obtained 157 articles with author's keywords information. The initial keyword analysis of these articles yielded a total of 712 keywords. To ensure relevance and focus on core concepts, only keywords appearing in at least two distinct articles were retained for subsequent analysis. A final refined set of 26 keywords (excluding "*phygital*") was kept. These 26 keywords and the "*physical*" keyword were then ingested into G2 to create the final KM (presented in Figure 1). Several key relationships emerge:

Phygital as a Bridge: A primary function of *"phygital"* as depicted, is to bridge the digital and physical realms. This bridging role is explicitly linked to "simulation", "digital twin", "training" and "medical education." Phygital environments enable enhanced training and educational experiences by integrating digital representations and simulations with tangible, physical elements. This is further reinforced by the connection to "computational modelling" indicating the use of digital models to inform and interact with physical processes within a phygital context.

Phygital in Healthcare: The map positions "phygital" firmly within the domain of "healthcare" and, more specifically, within "Phygital Healthcare". This sub-concept highlights the application of phygital principles to healthcare practices. Relationships with "digital health", "telehealth ecosystem" and "remote patient monitoring (RPM)" indicates that phygital approaches are integral to the evolving landscape of digital healthcare delivery.

Technological Enablers of Phygital: Several technological concepts are shown to enable or contribute to the creation and function of phygital environments. "Technology" itself forms a broad category, encompassing "artificial intelligence", "machine learning" and "deep learning". These AI-related technologies are shown to power "AI-Powered Diagnostics" and contribute to "precision medicine", suggesting that AI plays a critical role in analyzing data and personalizing treatments within phygital healthcare settings. Furthermore, "3D printing" is represented as enabling the "physical manifestation of digital designs", a core component of creating tangible elements within a *phygital* environment. The connection to "virtual reality" and "metaverse" indicates that immersive digital environments are also key components of certain *phygital* applications, potentially offering virtual spaces for training, simulation, or patient interaction.

Contextual Factors: contextual factors influencing the development and adoption of *phygital* approaches are present in the map. "COVID-19" is linked to "accelerated adoption of digital solutions", suggesting that the pandemic has spurred the development and implementation of phygital technologies in healthcare.

In conclusion, the generated KM provides a valuable overview of the multifaceted nature of the *phygital* concept and its diverse applications within healthcare. It highlights the role of phygital as a bridge between the digital and physical worlds, emphasizing its potential to enhance training, education, and healthcare delivery. The map also underscores the importance of enabling technologies such as AI, 3D printing, and virtual reality in creating effective *phygital* environments. Withing this context, the Rome technopole project appears to be fully aligned with the guidance provided by KM. The proposed innovative methodology answered the main research question of this article, laying the foundation for a first AI KM creation grounded in bibliometric metadata. Although this methodology shows interesting results, limitations must be acknowledged. The first concerns the use of the G2 tools, which, despite being one of the most advanced AIs, still has biases deriving from its training

process and available information. The further limitation concerns the prompt generated using *termboard* can be improved by using advanced prompting techniques such as few shots (Kaplan et al., 2020) or chain of though (Wei et al., 2022).

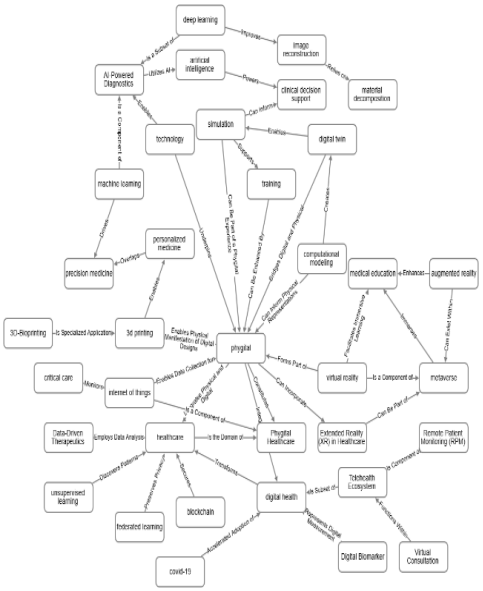


Figure 1. AI knowledge map on phygital in the healthcare sector, grounded on bibliometric metadata.

Acknowledgments

The poster is part of the Research Projects "Phygital Twin Technologies for Innovative Surgical Training & Planning (ECS 0000024 – Rome Technopole)".

References

Batat, W. (2024). What does phygital really mean? A conceptual introduction to the phygital customer experience (PH-CX) framework. Journal of Strategic Marketing, 32(8), 1220–1243.

Haustein, S., & Larivière, V. (2014). The use of bibliometrics for assessing research: Possibilities, limitations and adverse effects. In Incentives and performance: Governance of research organizations (pp. 121–139). Springer.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. ArXiv Preprint ArXiv:2001.08361.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V, & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

https://doi.org/10.51408/issi2025_198

India and quantum computing-related publications

Xiaojun Hu[1], Ronald Rousseau[2]

[1] *xjhu@zju.edu.cn*
Medical Information Center, Zhejiang University School of Medicine, Hangzhou 310058 (China)

[2] *ronald.rousseau@uantwerpen.be, ronald.rousseau@kuleuven.be*
Faculty of Social Sciences, University of Antwerp, Antwerp, 2020 (Belgium)
Facultair Onderzoekscentrum ECOOM, MSI, KU Leuven, Naamsestraat 61, Leuven, 3000 (Belgium)

Introduction

We explore how India performs in the field of quantum computing. An Indian publication is defined as one in which at least one author has an Indian address. This investigation has the following purposes:

(1) To perform a bibliometric study of India as an upcoming country in the field of quantum computing.

(2) To perform a diachronous study of its yearly citations.

(3) To investigate the influence of collaborations with the USA on India in the field of quantum computing

Methodology

Using the Web of Science (WoS), we consider the period [2003, 2023] and focus on Indian publications as defined above. We used the following query: ALL = (qubit* OR (quantum AND (comput* OR algorithm* OR crypto* OR gate* OR superposition* OR complexit*))), Although we also considered (TS=)-queries, we prefer (All=)-queries for this investigation. In this way, we can include in the resulting set, institutes, conferences, and funding sources with e.g., the phrase "quantum computation" in their name. The difference between (TS=) and (ALL=) queries reflects the importance of "quantum computing" in the general scientific landscape in the country under consideration. For example, when a university has a department with quantum computing in its name, or when "quantum computing"- funds are available, this is a sign of the importance given to quantum computing in the country and its collaborating countries.

Results: publications

Total number of publications, (ALL=) versus (TS=) queries

In the whole database, we found 14,170 Indian articles (September 2024), where we have included here all publications (of article or review type) with at least one Indian co-author. Table 1 shows the number of publications. The resulting data are diachronous ones as they show how data change over time. The curves for (ALL=) and for (TS=) queries both show an exponential increase.

Table 1. Indian publications: results for (ALL=) and (TS=) queries, and their ratio.

Year	ALL	TS	%	Year	ALL	TS	%
2003	88	66	0.75	2014	448	321	0.72
2004	85	82	0.96	2015	518	374	0.72
2005	102	85	0.83	2016	565	389	0.69
2006	132	109	0.83	2017	628	412	0.66
2007	119	96	0.81	2018	769	542	0.71
2008	159	129	0.81	2019	847	581	0.69
2009	200	153	0.77	2020	1075	716	0.67
2010	222	181	0.82	2021	1449	920	0.64
2011	280	213	0.76	2022	1641	1071	0.65
2012	303	229	0.76	2023	1875	1187	0.63
2013	385	286	0.74				

The (TS/ALL)-ratio starts at about 0.81-0.86 in the early years and reaches a low of 0.63 in 2023.

Domestic production

In this section, we study the percentage of domestic production (all authors have an Indian address) in the total of the country. Results are shown in Table 2. The percentage of domestic production has a decreasing tendency. Note that all non-domestic publications are publications resulting from international collaboration. The increase of internationally collaborated publications by India in the field of quantum computing corresponds with the growth of international collaboration in general.

Table 2. Percentages of domestic production for India.

Year	Percentage	Year	Percentage
2003	0.75	2014	0.70
2004	0.81	2015	0.71
2005	0.73	2016	0.72
2006	0.78	2017	0.64
2007	0.74	2018	0.67
2008	0.75	2019	0.67
2009	0.76	2020	0.64
2010	0.72	2021	0.64
2011	0.68	2022	0.60
2012	0.73	2023	0.61
2013	0.72		

India-US collaborated publications

Next, we study the number of collaborated articles and their evolution between the USA and India. We chose the USA as this country was until recently (when China took over) the leading country in the field of quantum computing. By the term "collaborated article" we mean an article with addresses of two countries under study. Other countries may participate in such articles.

In absolute numbers the collaboration between the USA and India in quantum computing is increasing. This observation also holds relatively speaking (over the period [2003, 2023]): India participated in less than 1% of US quantum publications and increased its share to more than 6%, while the US had a share of almost 10% of India's publications, increasing to more than 16%. These numbers illustrate the unequal scientific relationship between India and the USA.

Results: Citations

We counted (in the WoS) the number of received citations for each publication year in the period 2003-2013. These citations were counted over a ten-year period (plus the publication year) if possible, i.e., for the publication years 2003-2013, and over 5-years (plus the publication year) for the period 2003-2018. For each publication year, we also determined the h-index again over a ten-year period, and over a 5-year period. Similarly, we determined the average number of citations and the median number of citations. We did this for all publications, domestic publications only, and for the collaboration with the USA.

Citations of all Indian publications

Table 3 shows the data for all Indian publications. The five columns refer respectively to the publication year (FPY), the sum of all citations received by all publications (of the year in the first column) over a ten-year period (plus the publication year) (CIT10), the h-index calculated over the ten-year period (h10), the average number of received citations over the ten-year period (AV10) and finally the median number of received citations over the ten-year (MED10).

Table 3. Citations received by all Indian publications.

FPY	CIT10	h10	AV10	MED10
2003	1,115	16	12.67	5
2004	1,380	21	16.24	8
2005	1,847	22	18.11	7
2006	1,724	21	13.06	7
2007	2,114	21	17.76	9
2008	3,830	32	24.09	10
2009	4,564	33	22.82	13
2010	4,241	31	19.1	11
2011	5,953	37	21.26	10
2012	6,804	43	22.46	13
2013	9,941	46	25.82	14

There is a clear linear increase for the h-indices, the average number of received citations and the median number of citations. As h-indices depend on the number of publications and the number of citations and as both are increasing, it is obvious that also the h-index should increase.

Citations of India-US collaborated papers

In this section we present citation data for the US-India collaborations, see Table 4. The numbers of India-American joint publications are relatively small and rather irregular. Data refer to a 5-year citation window.

Table 4. Citation data for US-India collaborated papers.

FPY	# PUB	CIT5	h5	AV5	MED5
2014	55	2,734	23	49.7	18
2015	55	1,880	23	34.2	17
2016	66	1,828	21	27.7	13
2017	89	4,236	22	47.6	14
2018	114	3,155	26	27.7	15

Conclusion

We have studied all quantum computing-related articles and reviews in the WoS with at least one Indian author in the period [2003, 2023]. We separately considered domestic publications and collaborations with the USA. Besides the number of publications we also considered yearly citations and the corresponding h-indices. There is a generally increasing trend for all indicators, but a relative decrease in domestic publications.

References

Fassin, Y., & Rousseau, R. (2022). On the difference between an (ALL=)- and a (TS=)-query in the Web of Science: the case of bibliometrics versus scientometrics. *ISSI Newsletter*, 18(2), #70, 30-33..

https://doi.org/10.51408/issi2025_176

Machine learning-based model to predict topics contributing to Sustainable Development Goals: A study of Latin American and European Countries

Barbara S. Lancho Barrantes

b.lanchobarrantes@brighton.ac.uk
School of Architecture, Technology and Engineering, University of Brighton
Brighton BN2 4GJ (United Kingdom)

Introduction

Health is a human right and a cornerstone of physical, mental, and social well-being (WHO, 1949). Ensuring access to healthcare is not only ethical; it is essential for reducing poverty and fostering inclusive, sustainable development.

Yet, health systems worldwide face persistent challenges: underfunding, high out-of-pocket costs, fragmented service delivery, and inequities based on the ability to pay. These issues, compounded by ineffective governance, undermine progress toward universal health coverage and financial protection.

The 2030 Agenda for Sustainable Development, adopted by the United Nations in 2015, includes 17 SDGs, with Goal 3 dedicated to ensuring healthy lives and well-being for all. Though interlinked with other goals, SDG 3 plays a pivotal role in shaping global health priorities.

Recent bibliometric studies have explored how countries' research aligns with SDG challenges (Yamaguchi et al. 2023)However, much of this work focuses on sectors like business or education, leaving a gap in understanding SDG 3-related research, especially in Global South and Global North countries (Yaqub, et al., 2024) Notably, low-income countries, despite facing the greatest SDG-related challenges, contribute minimally to the research driving global progress (Confraria et al., 2024).

This study uses the OpenAlex (Priem et al.2022) database to examine how countries' SDG 3 research priorities differ. It highlights the need for context-sensitive bibliometric strategies that account for regional health challenges often overlooked in global analyses.

By analysing large-scale publication data, research trends, topics, and collaboration patterns, AI tools can predict emerging health research themes with growing accuracy. The research questions of this study are:

- How do research priorities differ between Latin America (Global South) and Europe (Global North)?
- Do both regions focus on similar health challenges with equal intensity?
- Can machine learning-driven models effectively forecast future research trends?

Data and Methods

This study draws on data from OpenAlex, a large open-access bibliographic database with over 240 million scholarly works. The study focused on SDG 3: Good Health and Well-being, selecting the top 10 publishing countries from:

- Latin America: Brazil, Uruguay, Mexico, Colombia, Chile, Costa Rica, Puerto Rico, Argentina, Ecuador, and Peru.
- Europe: United Kingdom, France, Germany, Spain, Netherlands, Switzerland, Italy, Belgium, Sweden, and Poland.

OpenAlex leverages a machine learning-based SDG Classifier to assess the relevance of academic publications to the 17 Sustainable Development Goals (SDGs). Using Natural Language Processing (NLP), the system analyses titles, abstracts, keywords, and citations to understand the content of each

publication. The SDG BERT model, a multilingual, multi-label transformer trained on SDG-labeled data (Aurora Query Model v5), then assigns a probability score (ranging from 0 to 1) for each SDG, indicating the publication's relevance.

In addition to SDG tagging, OpenAlex employs an automated topic classification system that assigns each publication to one or more of ~4,500 scientific topics.

Results

Table 1 compares health research priorities and contributions to SDG 3 between Latin American and European countries, highlighting differences in focus and scientific output.

Publications count	Classification		
Topic	Europe	Latin America	Total
SARS-CoV-2 and COVID-19 Research	4954	953	5907
Cancer Immunotherapy and Biomarkers	3931	300	4231
COVID-19 and Mental Health	2678	1175	3853
Liver Disease Diagnosis and Treatment	3265	581	3846
Cardiac Valve Diseases and Treatments	3435	231	3666
COVID-19 Clinical Research Studies	2510	1038	3548
Inflammatory Bowel Disease	2981	220	3201
Atrial Fibrillation Management and Outcomes	2844	200	3044
Diabetes Treatment and Management	2655	326	2981
CAR-T cell therapy research	2919	9	2928
Long-Term Effects of COVID-19	2147	765	2912
Mosquito-borne diseases and control	1291	1353	2644
Pancreatic and Hepatic Oncology Research	2402	239	2641
Prostate Cancer Treatment and Research	2348	196	2544
COVID-19 and healthcare impacts	2020	491	2511
Lung Cancer Treatments and Mutations	2306	137	2443
Acute Myeloid Leukemia Research	2266	177	2443
Tuberculosis Research and Epidemiology	1848	592	2440
Glioma Diagnosis and Treatment	2132	232	2364
Acute Ischemic Stroke Management	1968	362	2330

Europe leads in publications on advanced medical topics like cancer and cardiac diseases, while Latin America focuses on mosquito-borne diseases, mental health, and long-term pandemic effects. COVID-19 research is a shared focus, with Europe concentrating on clinical studies and Latin America on mental and social health impacts. These differences reflect regional health priorities and socioeconomic factors.

A Random Forest classifier is used to predict publication origin based on research topics. The model's accuracy improved with more data but was affected by class imbalance, as Europe had more publications, potentially introducing bias.

The model shows a moderate ability to distinguish between regional research focuses, but performance is limited. Despite a balanced dataset (Europe: 422, Latin America: 378), topic overlap likely reduced prediction clarity. Results suggest that thematic differences exist but are not strongly distinctive based on topic data alone.

```
Model: Logistic Regression
Accuracy: 0.55625
Classification Report:
              precision    recall  f1-score   support

      Europe       0.59      0.53      0.56       422
Latin America      0.53      0.58      0.55       378

    accuracy                           0.56       800
   macro avg       0.56      0.56      0.56       800
weighted avg       0.56      0.56      0.56       800

------------------------------------------------------------
Model: Random Forest
Accuracy: 0.575
Classification Report:
              precision    recall  f1-score   support

      Europe       0.60      0.58      0.59       422
Latin America      0.55      0.57      0.56       378

    accuracy                           0.57       800
   macro avg       0.57      0.57      0.57       800
weighted avg       0.58      0.57      0.58       800

------------------------------------------------------------
Model: Naive Bayes
Accuracy: 0.55125
Classification Report:
              precision    recall  f1-score   support

      Europe       0.59      0.48      0.53       422
Latin America      0.52      0.63      0.57       378

    accuracy                           0.55       800
   macro avg       0.56      0.56      0.55       800
weighted avg       0.56      0.55      0.55       800

------------------------------------------------------------
```

Future research should focus on individual countries' contributions to SDG 3 rather than on regions. Analysing publication abstracts and exploring models like Gradient Boosting (e.g., XGBoost, LightGBM) and deep learning-based NLP could reveal subtle patterns. Additionally, optimising the Random Forest model with techniques like Grid or Random Search could enhance performance.

Conclusions

The number of publications on SDG 3: Good Health and Well-Being has increased significantly in the last two years, creating challenges for researchers to identify priorities among countries. This study used machine learning and NLP to track shifts in health research topics, from "SARS-CoV-2" to emerging areas like "Zika" and "Cancer." It

highlights gaps in research on topics like "Palliative Care" and "Cerebral Venous Sinus Thrombosis." By analysing open research data, the study predicted future trends for 10 Latin American and 10 European countries, revealing ongoing regional differences in health priorities. It emphasises the need for stronger partnerships, more funding, and improved capacity in Latin America. Overall, machine learning and NLP enhance research efficiency and support decision-making for SDG 3.

References

Confraria, H, Ciarli, T & Noyons, E, (2024). Countries' research priorities in relation to the Sustainable Development Goals. Research Policy, Elsevier, vol. 53(3).

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833

Yaqub, O, Coburn, J & Moore, D A.Q. (2024). Research-targeting, spillovers, and the direction of science: Evidence from HIV research-funding. Research Policy, 53(8).

Yamaguchi, N. U., Bernardino, E. G., Ferreira, M. E. C., de Lima, B. P., Pascotini, M. R., & Yamaguchi, M. U. (2023). Sustainable development goals: A bibliometric analysis of literature reviews. Environmental Science and Pollution Research, 30(3), 5502–5515.

https://doi.org/10.51408/issi2025_179

Measuring the effect of research award on collaboration relationships

Chien Hsiang Liao

jeffen@gmail.com
Fu Jen Catholic University, Department of Information Management, No. 510 Zhongzheng Rd., Xinzhuang Dist., New Taipei City, 24205 (Taiwan)

Introduction

A scholar's reputation carries significant weight. Scholars with outstanding research performance (e.g., Nobel Prize winners) tend to receive more attention and attract other scholars to cite their articles (Inhaber and Przednowek, 1976). A scholar's reputation derived from research awards possess 'halo effect', implying that human's perception is easily affected by a given impression. Reputation not only brings more external resources for the awardees, but also helps expand academic social networks (Li et al., 2013). A good reputation may contribute to expanding social capital (social network), attracting more scholars' attention and expanding the collaboration network.

More specifically, Liao (2021) indicates that cumulative advantages of research funding in the past and institutional reputation positively affect the amount of future research funding. Nevertheless, although the impacts of these effects on future research funding have been proven, the causal relationship between effects has not been fully revealed. In the existing literature, there is little mention of how the halo effect affects social influences. Therefore, this study is initiated to explore the differences in the social capital (collaboration relationships) at different periods before and after receiving research awards (i.e., halo effect). The measures of collaboration relationships include the numbers of (1) nonrepeated collaborators, (2) international research articles, and (3) cross-disciplinary research collaboration articles. The research targets are the awardees of the National Science and Technology Council (NSTC) in Taiwan from 2011 to 2017. The specific research questions are listed as below.

RQ: Does research award (the halo effect) breed the social capital?

Methods

Data sources

There are two sources of secondary data in this study. The first source is the website of the NSTC in Taiwan, where you can query the information of the NSTC research awards, including awardees, years and their affiliations. In addition, on this website, this study collects the research project information of all awardees, including the project name, applicant information, approved funding, and project execution duration. However, due to the complexity and wide range of disciplines, the research performance and funding done by applicants at the different disciplines are very distinct. Considering the possible influences of different disciplines (external variables), this study limited the data to the 'Management Science' discipline. For example, NSTC works of the awardees in the discipline of Humanities are specific books or exhibitions, and it is difficult to compare outcomes with other disciplines. More specifically, the data of this study focus on the winners of the NSTC research awards in the 'Management Science' discipline from 2011 to 2017. This period was chosen because of the need to compare research performance over times. According to the winner's award year, this study collects (a) the data of four-years research projects before they awarded, the data interval is from 2008 to 2016; and (b) the data of four-years research projects after they awarded, the data interval is from 2013 to 2021.

The second data source is the NSTC talent database which documents the research performance of scholars. All scholars must be registered in this talent database to apply for

the NSTC projects, so the database also contains research publications of all awardees. At this database, this study collects their publications before and after they awarded. The data include the amount of journal article.

Measures

About the measurement of the proposed effects, the NSTC research award is treated as the manifestation of the 'halo effect', reflecting whether the awardee can enjoy the advantages of reputation after winning the award. The NSTC research award is the NSTC credible research award in Taiwan, only a few winners are awarded each year. The NSTC only grants two types of research awards, including Outstanding Research Award and Ta-You Wu memorial award for young talent under the age of 42 (Liao, 2021). To carefully investigate the halo effect, this study will also analyse these two awards separately.

Regarding 'social capital', the concept measures whether the awardees' collaboration network expands or not, including calculating (1) the number of nonrepeated collaborators in their publications before and after they are rewarded. Among their publications, if there are two or more coauthors from different countries in an article, it will be calculated as an (2) international collaboration article. Likewise, if there are more coauthors in different disciplines, it will be calculated as a (3) cross-disciplinary collaboration article. The determination of international and cross-disciplinary collaboration is based on the information of coauthors nationality, academic institution or affiliation in the article. This study focuses on awardees in the field of Management Science, most of whom have published SSCI journal articles. These articles are accessible through Google Scholar or Web of Science, which provide information on the authors' affiliations and countries. Discipline classifications follow the category definitions in the Journal Citation Reports. If a co-author's institution falls outside the Management Science field, the collaboration is classified as cross-disciplinary.

The measurement and interval of indicators is showed in Figure 1.

Research indicators	Time interval						
Halo effect							
- The MOST research award (year)	2011	2012	2013	2014	2015	2016	2017
Before the award							
Social capital							
- N of nonrepeated collaborators	2007-2010	2008-2011	2009-2012	2010-2013	2011-2014	2012-2015	2013-2016
- N of international collaborations	2007-2010	2008-2011	2009-2012	2010-2013	2011-2014	2012-2015	2013-2016
- N of cross-disciplinary collaborations	2007-2010	2008-2011	2009-2012	2010-2013	2011-2014	2012-2015	2013-2016
After the award							
Social capital							
- N of nonrepeated collaborators	2012-2015	2013-2016	2014-2017	2015-2018	2016-2019	2017-2020	2018-2021
- N of international collaborations	2012-2015	2013-2016	2014-2017	2015-2018	2016-2019	2017-2020	2018-2021
- N of cross-disciplinary collaborations	2012-2015	2013-2016	2014-2017	2015-2018	2016-2019	2017-2020	2018-2021

Figure 1. The measurement and interval of indicators.

Results and conclusions

The proposed associations are examined by using paired sample T-test and compare the difference between performances at the different periods. The results are illustrated in Table 2. For the impact on social capital, the results show that the number of articles on international collaboration (T = 2.044; P < .05) and cross-disciplinary collaboration (T = 2.012; P < .05) has significantly increased for all awardees. However, if the sample is divided into two award groups, the award (i.e., the Ta-You Wu Memorial Award) does not significantly help young talents to expand social capital and enhance their collaboration network. In contrast, after experienced scholars won the Outstanding Research Award, the number of collaborators (T = 1.742; P < .1), international collaborations (T = 2.022; P < .05), and cross-disciplinary collaborations (T = 2.066; P < .05) all positively increased. The proposed associations are partially supported, revealing that the halo effect will enhance social capital only for experienced scholars. A reasonable explanation is that experienced scholars have had more time to build up a broad network of contacts within the field, including other researchers, institutions, and organizations. Once these experienced scholars gain recognition through the award, the network effect will be higher than that of young scholars. These network effects may come from the direct or indirect influence of their relationships in the network, such as collaboration opportunities.

Paired sample	Paired difference			
	Mean	Std deviation	T value	Significance
All (n=122)				
N of nonrepeated collaborators (After – Before)	2.566	17.314	1.637	.104
N of international collaborations (After – Before)	.86	4.65	**2.044**	**0.43***
N of cross-disciplinary collaborations (After – Before)	1.279	7.02	**2.012**	**0.46***
Ta-You Wu Memorial Award – Young talent (n=43)				
N of nonrepeated collaborators (After – Before)	.103	15.968	.04	.968
N of international collaborations (After – Before)	.0513	3.178	3101	.920
N of cross-disciplinary collaborations (After – Before)	.0769	2.366	.203	.840
Outstanding research award (n=79)				
N of nonrepeated collaborators (After – Before)	3.582	18.278	**1.742**	**.085***
N of international collaborations (After – Before)	1.165	5.12	**2.022**	**0.47***
N of cross-disciplinary collaborations (After – Before)	1.975	8.497	**2.066**	**0.42***

Figure 2. The results of T-test.

The findings show that the halo effect only breeds social capital for experienced scholars. This research finding provides evidence for the importance of social capital in academic research. Experienced scholars who have had more time to build up a broad network of contacts are likely to have greater access to resources, information, and opportunities, which can facilitate their research activities and collaboration. Corresponding to the statement of the network effect, this study found that recognition through awards is not only a signal of individual excellence but also a reflection of the social embeddedness of scholarly work. For practical implications, this study suggests that young scholars may need more support to build up their own networks of contacts within their field. This could include mentoring programs, networking events, and opportunities for collaboration with experienced scholars and other professionals.

In conclusion, funding agencies need to consider the different needs and priorities of young and senior scholars when designing and allocating research awards. This study highlights the importance of social capital in academic research and suggests that building up networks of contacts within the field should be a priority for scholars at all stages of their careers.

References

Inhaber, H., & Przednowek, K. (1976). Quality of research and the Nobel prizes. *Social Studies of Science*, 6(1), 33-50.

Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9), 1515-1530.

Liao, C. H. (2021). The Matthew effect and the halo effect in research funding, *Journal of Informetrics*, 15(1), 101108.

https://doi.org/10.51408/issi2025_199

MetaInfoSci: Visualize trends and understand facts

Kiran Sharma[1], Parul Khurana[2], Ziya Uddin[3]

[1] *kiran.sharma@bmu.edu.in*, [3]*ziya.uddin@bmu.co.in*
School of Engineering & Technology, BML Munjal University, Gurugram, Haryana-122413 (India)
Center for Advanced Data and Computational Science, BML Munjal University, Gurugram, Haryana-122413 (India)

[2] *parul.khurana@lpu.co.in*
School of Computer Applications, Lovely Professional University, Phagwara, Punjab-144411 (India)

Introduction

The exponential growth of scientific literature has created an imperative need for effective analysis by understanding and analysing publication trends and to navigate the complex requirements of academics, research institutions, and policymakers (Jacob and Meek, 2013). The ability to gauge the growth of scientific disciplines, author collaboration, and institutional partnerships offers valuable insights that can inform research directions, funding decisions, and educational strategies (Bozeman, Fay and Slade, 2013). However, this requires users to possess technical expertise in data cleaning and formatting, which results in creating barriers to entry for researchers from diverse disciplines and their ability to comprehend the full scope of their fields.

Current bibliometric tools often fall short, providing quantitative outputs such as citation counts and co-authorship networks without the necessary interpretation to make these figures meaningful (Zupic and fater, 2015). This gap leaves researchers without a clear understanding of how their work contributes to broader scientific progress and a barrier for the adoption of these tools. Further, existing network platforms typically require users to input data in clean formats, which can be daunting for those lacking technical expertise. To bridge this gap, we propose the development of innovative MetaInfoSci, a comprehensive web tool designed to streamline both qualitative and quantitative analysis of research literature. It will not only automate data preprocessing and cleaning, making it accessible to non-technical users, but will also employ advanced AI algorithms to interpret and contextualise research findings.

Objective of the study

The MetaInfoSci web tool is designed as an all-in-one, user-friendly platform for conducting bibliometric, scientometric, and network analysis of bibliographic databases such as Scopus, WoS, and OpenAlex. It features real-time visualization updates and AI integration for result interpretation. Additionally, the tool will incorporate network science metrics, Gender API integration, Journal Quartile integration, and more.

Relevant Existing Tools

Table 1. shows major tools with their purpose, features, and limitations.

Tool Name	Purpose	Features	Limitations
VOSviewer, Van Eck and Waltman, 2017	Bibliometric visualization	Co-occurrence analysis, Bibliographic coupling, Network visualization, Clustering	Lacks advanced citation metrics, limited interoperability with other tools, steep learning curve for beginners

Gephi, Bastian, Heymann and Jacomy, 2009)	Network visualization	Graph visualization, Network clustering, Community detection, Dynamic graph analysis	Resource-intensive for large networks, lacks built-in bibliometric functions, steep learning curve
ScientoPy, Ruiz-Rosero, Ramírez-González and Viveros-Delgado, 2019	Bibliometric analysis	Trend analysis, Performance metrics, Bibliographic coupling, Statistical analysis	Limited visualization options, requires manual data preparation, lacks cloud-based accessibility
BibExcel, Persson, Danell and Schneider, 2009	Bibliometric analysis	Bibliographic data conversion, Excel output, Network preparation, Citation analysis	Not user-friendly for large datasets, lacks real-time updates, limited interactivity
Biblioshiny, Aria and Cuccurullo, 2017	Bibliometric analysis	Shiny-based GUI, Bibliometrix integration, User-friendly statistical visualization	Limited customization, requires R setup, lacks real-time bibliometric updates

Demonstration

Home Page

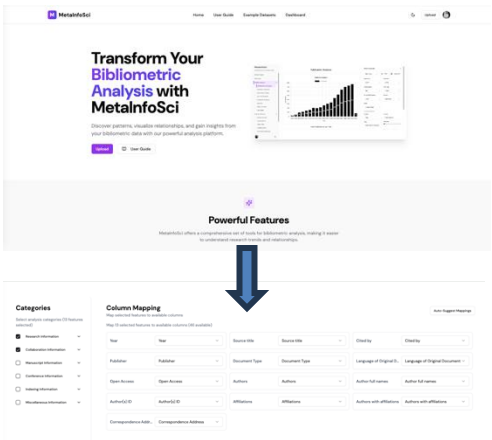


Figure 1. Uploading data and mapping columns.

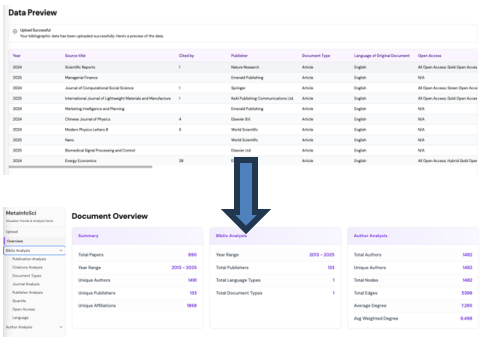


Figure 2. Displaying data and overview of data.

Publications Analysis

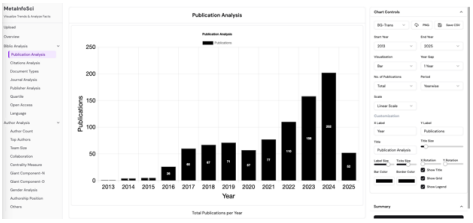


Figure 3. Under publication trends analysis tab, displaying year-wise trend analysis.

Journal Analysis

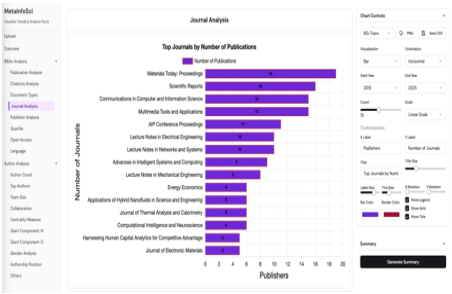


Figure 4. Under journal analysis tab, displaying bar plot of journal count.

Author's Analysis

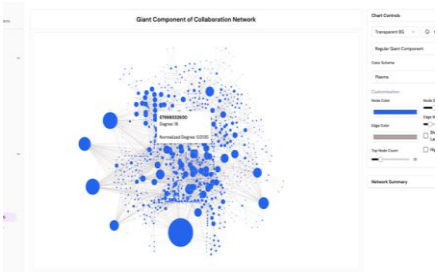


Figure 5. Under author's analysis tab, displaying authors 'collaboration network.

How it is different from existing tools?

1. Interactive visualization tool box
2. AI-integrated result interpretation
3. Advanced network metrics and analysis
4. Advanced bibliometrix analysis like journal quartile, gender analysis, etc.

Conclusion

The proposed study aims to develop MetaInfoSci, an advanced web tool for both qualitative and quantitative research literature analysis, integrating bibliometric, scientometric, and network analysis. Unlike existing platforms, MetaInfoSci will unify data from multiple sources such as Scopus, Web of Science, OpenAlex, etc. into a single dataset. The tool will offer features like publication trend analysis, identification of key contributors, and detailed collaboration networks at the author, institution, and country levels. Powered by AI-driven algorithms, MetaInfoSci will provide meaningful context and interpretation for bibliometric data, enabling researchers to better understand the broader impact of their work. Additionally, it will incorporate outer datasets on journal quality metrics (e.g., Quartiles, h-index, impact factor) and author gender, facilitating in-depth and customizable analysis with flexible visualizations. By automating data merging, cleaning, and analysis, MetaInfoSci will make bibliometric tools more accessible to non-technical users, while empowering researchers with advanced analytical capabilities. The platform will also serve as a valuable training resource, equipping users with essential skills for navigating the evolving landscape of research evaluation.

Acknowledgment

The author acknowledges R&D Cell for their financial support through the seed grant (No: BMU/RDC/SG/2024-06).

References

Aria, M., Cuccurullo, C., 2017. bibliometrix: An r-tool for comprehensive science mapping analysis. Journal of informetrics 11, 959–975.

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks, in: Proceedings of the international AAAI conference on web and social media, pp. 361–362.

Bozeman, B., Fay, D., Slade, C.P., 2013. Research collaboration in universities and academic entrepreneurship: the-state-of-the-art. The journal of technology transfer 38, 1–67.

Jacob, M., Meek, V.L., 2013. Scientific mobility and international research networks: trends and policy tools for promoting research excellence and capacity building. Studies in higher education 38, 331–344.

Persson, O., Danell, R., Schneider, J.W., 2009. How to use bibexcel for various types of bibliometric analysis. Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday 5, 9–24.

Ruiz-Rosero, J., Ramírez-González, G., Viveros-Delgado, J., 2019. Software survey: Scientopy, a scientometric tool for topics trend analysis in scientific publications. Scientometrics 121, 1165–1188.

Van Eck, N.J., Waltman, L., 2017. Citation-based clustering of publications using citnetexplorer and vosviewer. Scientometrics 111, 1053–1070.

https://doi.org/10.51408/issi2025_205

Multidimensional quantitative analysis of the fit of Chinese science and technology talent policy

Wang Kaile[1], Chen Yunwei[2]

[1]*wangkl@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, No.289, QunXian Nanjie, Tianfu New Area, Chengdu (China)

[2]*chenyw@clas.ac.cn*
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, No. 1 Yanqi Lake East Road, Huairou District, Beijing (China)

Introduction

High-quality science and technology talent policies are essential for unlocking the innovative potential of scientific and technological talent, addressing the mismatches between talent supply and the demand for technological innovation, advancing the national strategy for science and technology talent, and fostering a conducive environment for talent development. In this context, deconstructing the policy framework for China's scientific and technological talent, analysing gaps, redundancies, and deficiencies in policy design, and optimizing the policy system are critical and timely endeavours.

This paper proposes a multidimensional analytical framework for assessing the fit of science and technology talent policies. It seeks to interpret the characteristics and challenges of China's science and technology talent policies through three dimensions: consistency fit, embedded fit, and compensatory fit. Specifically, this framework examines: the internal coordination of policies (consistency fit), the compatibility between policies and the overarching policy system (embedded fit), and the complementarity between policy supply and talent demand (compensatory fit). By employing quantitative analysis methods, this study aims to contribute to the theoretical research on talent policies, provide methodological insights, and offer practical recommendations for optimizing China's science and technology talent policy system.

Materials and methods

The analysis of policy fit comprises two components: the policy analysis module and the talent analysis module. The policy analysis module primarily examines two sub-dimensions: policy supply and policy goals, while the talent analysis module focuses on two sub-dimensions: talent demand and talent goals (Figure 1). These four sub-dimensions form six key interrelationships: the relationship between policy supply and policy goals—If the tools provided by the policy (policy supply) align effectively with its stated objectives (policy goals), consistency fit is achieved; the relationship between policy supply and talent goals—If the policy integrates seamlessly into the broader policy framework and serves as an effective tool for achieving talent goals, embedded fit is established; the relationship between policy supply and talent demand—If the policy supply meets the specific needs of talent (talent demand), compensatory fit is attained; the relationship between policy goals and talent goals—Since policies are inherently designed to achieve talent development goals, there is an automatic fit between these two dimensions; the relationship between policy goals and talent demand—The primary aim of a policy is to address talent demand, which serves as a sufficient condition for policy goals. Thus, an automatic fit exists between these two dimensions; and the relationship between talent demand and talent goals—Talent goals are derived from individual and collective talent needs, with collective needs often

reflecting government requirements for talent development. Talent goals are therefore a necessary condition for talent demand, creating an automatic fit between these dimensions.

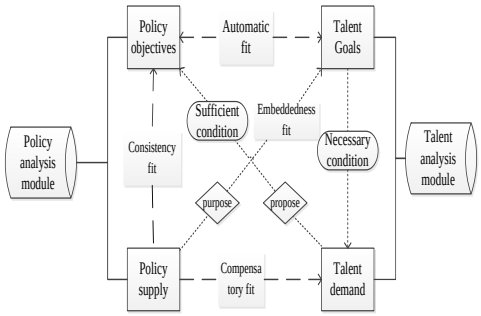


Figure 1. Relationship diagram for analysing the fit of talent policies.

Empirical Analysis of the Fit Degree of China's Talent Policies

After a decade of development, China's science and technology talent policy system has essentially established a relatively stable framework. Both the consistency fit and compensatory fit values of the policies exhibit an upward trend, reflecting an improvement in the quality of China's science and technology talent policies. However, certain shortcomings persist in the formulation and implementation of these policies, as detailed below:

Limited Use of Development Planning Policy Tools: The utilization of development planning policy tools remains relatively limited, resulting in a lack of comprehensive top-level design for talent development. Instead, fragmented and piecemeal policy measures are frequently employed. This shortfall is one of the key reasons why the goal of improving talent quality in China has not been effectively achieved.

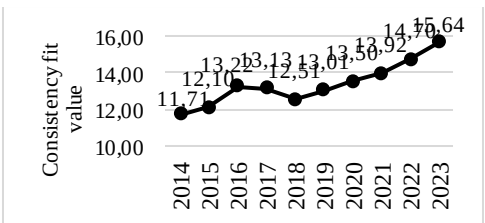


Figure 2. Consistency fit value of China's science and technology talent policy.

Discrepancy in Policy Tool Usage: A significant imbalance exists in the application of policy tools, with training and development tools being used more frequently, while introduction and aggregation policies, as well as development planning policies, are comparatively underutilized. The insufficient use of introduction and aggregation policy tools is a major factor hindering the realization of the goal to expand the talent pool as envisioned in China's talent policies. Furthermore, the content of existing introduction policies reveals a stronger focus on the recruitment of domestic talent, with less emphasis on attracting international talent. While there is substantial attention given to attracting talent for innovation and entrepreneurship, the introduction of high-level talent remains a lower priority.

Gap Between Policy Goals and Actual Outcomes: A noticeable gap exists between the stated goals of policy implementation and their actual effects. For example, while China's science and technology talent policies aim to enhance talent quality, the outcomes have been suboptimal, indicating that the measures and intensity of tools designed to improve talent quality require further strengthening.

Mismatch Between Policy Supply and Talent Demand: Although China has implemented numerous policies to support talent development, data indicates that a gap persists between policy supply and talent demand. This issue warrants significant attention in the future development of China's science and technology talent policies. As the scale and capabilities of the talent pool continue to grow, the nature of talent demand is also evolving. On one hand, the coordinated use of multiple policy tools should be prioritized to achieve the goals of expanding the talent pool and enhancing talent quality. On the other hand, a detailed analysis of talent demand should be conducted. Strengthening protections for talent in areas such as knowledge, living conditions, services, and institutional support is essential for transitioning from a general talent base to a higher-quality talent pool.

Table 1. Evolution of compensatory fit value in China's science and technology talent policy.

Year	*Value of compensatory fit*	*Year*	*Value of compensatory fit*
2014	(0.707，1)	2019	(0.902，1)
2015	(0.767，1)	2020	(0.911，1)
2016	(0.774，1)	2021	(0.944，1)
2017	(0.821，1)	2022	(0.992，1)
2018	(0.842，1)	2023	(0.999，1)

References

Liu, R.J., Wang, J., Tang, L.J., et al. (2023). Quantitative analysis of maternal and child health talents policy in the Yangtze River Delta from the perspective of policy tools. *Chinese Journal of Health Policy*, 16(11), 31-38.

Barry, RA. (2024). Challenges achieving horizontal coherence across health and public security policies in formulating Uruguay's cannabis regulation. *Health Promotion International*, 39(5), daae136.

Qiu, Y.M., Shi, C. (2023). The operational logic of China's policy diffusion from an intergovernmental relations perspective——a case study of urban talent attraction policies. *Jiangxi Social Sciences*, 43(10), 183-191+208.

Chen, Q.Y, Ye, Y., Li, X.P. (2024). Quantitative evaluation and spatiotemporal evolution of China's regional talent policies. *Jiangsu Social Sciences*, 01, 166-175.

Mea, M, Newton, A, Uyarra, MC, et al. (2016). From Science to Policy and Society: Enhancing the Effectiveness of Communication. *Frontiers in Marine Science*,3,168.

Zimmermann, M, Pye, S. (2018). Inequality in energy and climate policies: Assessing distributional impact consideration in UK policy appraisal. *Energy Policy*, 123, 594-601.

Wang, H.J. (2023). A Study on the path of enhancing the entrepreneurial ability of science and technology talents in agriculture related digital business, based on the perspectives of policy compliance, resource integration, and skill expansion. *Management of Agricultural Science and Technology*, 42(06), 56-60.

Muchinsky P M, Monahan C J. (1987). What is person-environment congruence? Supplementary versus complementary models of fit. *Journal of Vocational Behavior*, 31(3), 268-277.

https://doi.org/10.51408/issi2025_192

Multilevel Structures, Connection and Balance: The Evolution of the Structure of Science

Yuxian Liu[1], Hongrui Yang[2], Ronald Rousseau[3], Raf Guns[4], Sisi Li[5], Yafang Fan[6], Helan Wu[7], Sanfa Cai[8]

[1] *yxliu@tongji.edu.cn*, [2] *2330024@tongji.edu.cn*, [5] *2231429@tongji.edu.cn*, [8] *csf@tongji.edu.cn*
Tongji University, Institute of Higher Education, Siping Road 1239, 200092 Shanghai (China)

[3]*ronald.rousseau@kuleuven.be*
University of Antwerp, Faculty of Social Sciences (Belgium)
Department of MSI, Centre for R&D Monitoring (ECOOM) (Belgium)

[4] *raf.guns@uantwerpen.be*
University of Antwerp, Centre for R&D Monitoring (ECOOM) (Belgium)
University of Antwerp, University Library (Belgium)

[6]*sonyafan@ustc.edu.cn*
University of Science and Technology of China, University of Science and Technology of China Library, Jinzhai Road 96, 230026 Hefei (China)

[7]*08052@tongji.edu.cn*
Tongji University, School of Physics Science & Engineering, Siping Road 1239, 200092 Shanghai (China)

Introduction

Efforts to map the structure of science began in the sixties with the work of Garfield, Sher, and Torpie (1964), among others. Since then, various other approaches have been developed. In this study, we examine how updates to the Web of Science (WoS) categories influence these scientific maps

Journal categories and their groups and broad categories

When categories in the WoS are updated, we wonder what influence this has on the resulting maps. In this contribution, we make a comparative study to answer this question. We collected data, using the same method as in Liu (2018) to construct a map of the structure of science of 2024. When exploring the logic and landscape of the knowledge system, multilevel structures are often used to map the structure of science (Li, 2016).

There are two subject categorization schemes provided by the WoS. One scheme is for the Journal Citation Reports (JCR) specifically. In this scheme, the journals in the JCR are assigned to categories. In the 2024 version of the JCR, we notice that the categories are further divided into 21 groups. Another scheme is shared by all Web of Science product databases. In this scheme, the objects of all the databases are divided into different research areas. Research areas are classified into five broad categories: Arts & Humanities; Life Sciences & Biomedicine; Physical Sciences; Social Sciences; and Technology. We add multidisciplinary as the sixth broad category.

Table 1 shows the basic framework used here to analyze the structure of science, while the resulting structures of science are shown in Figures 1a and 1b, using VOSViewer.

Results

JCR categories schemes and research areas are two subject categorization schemes provided by the WoS. We map the two schemes and obtain a multilevel structure of journals, JCR categories, JCR groups, and Broad Categories. The change of categories leads to a change in the structure of science. The structure based on the 2016 data is like two opposite poles, with Science and Technology at one pole, and

Humanities & Social Sciences at the other one. The categories in the structure of 2024 are connected and have a triangular shape. The first one is that art & humanity and social sciences are split into two clusters in 2024's structure. One cluster includes more of the categories of art and humanity, the other contains more of the categories of the social sciences. Most of the moved categories are in the broad categories of technology, life science & biomedicine. Changes are explained by the fact that knowledge itself evolves and has no clear borders between disciplines. It percolates through a multilevel structure as shown in this article.

Table 1. Multilevel subject categorization scheme used in this study.

Broad categories	*JCR groups*	*Number of JCR categories*	*Number of journals*
Art & Humanity	Arts & Humanities, Interdisciplinary	8	1016
	Philosophy & Religion	7	988
	Literature & Language	17	1628
	Visual & Performing Arts	10	930
Social Sciences	History & Archaeology	9	1403
	Psychiatry/Psychology	16	1555
	Economics & Business	21	3464
	Social Sciences, General	41	6561
Life Sciences & Biomedicine	Agricultural Sciences	7	441
	Biology & Biochemistry	34	4026
	Plant & Animal Science	17	1635
	Clinical Medicine	59	7627
Physical Sciences	Chemistry	21	2412
	Physics	34	3067
	Mathematics	12	1807
	Environment/Ecology	13	1753
	Geosciences	14	1112
Technology	Materials Science	17	1660
	Engineering	41	3663
	Computer Science	14	1619
Multidisciplinary	Multidisciplinary	36	5859

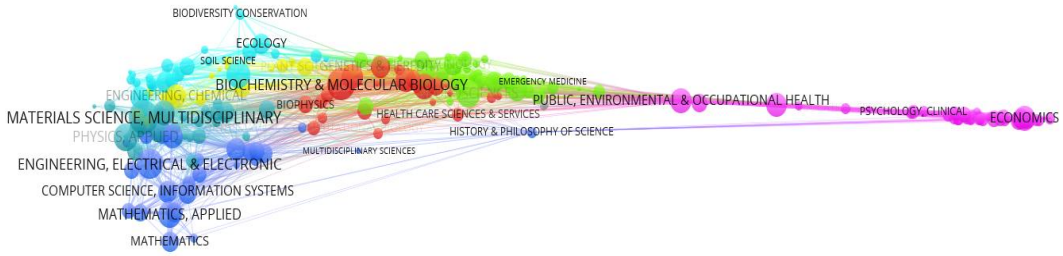


Figure 1a. 2016's structure of science constructed with JCR categories with the number of common journals as linkage strength.

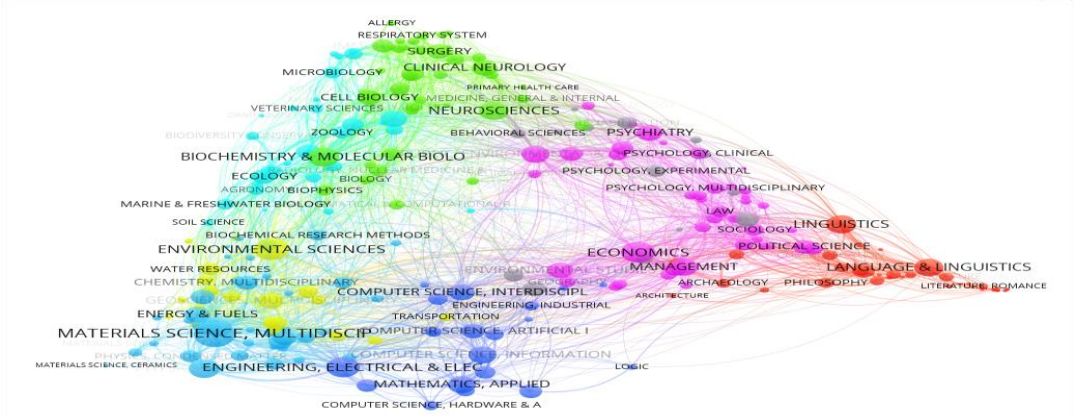


Figure 1b. 2024's structure of science constructed with JCR categories using the number of common journals as the linkage strength.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.72274139).

References

Li. J-H. (2016). Exploring the logic and landscape of the knowledge system: Multilevel structures, each multiscaled with complexity at the mesoscale. *Engineering*, 2(3), 276-285.

Liu, YX. (2018). Constructing a global backbone of science based on inter-categories co-membership of journals, *Journal of the China Society for Scientific and Technical Information*. 37(6), 580-589.

https://doi.org/10.51408/issi2025_207

Open Citations in German Educational Research–Identifying Disciplinary Practices to Train Data Extraction

Verena Weimer[1], Tamara Heck[2], Christoph Schindler[3]

[1]*v.weimer@dipf.de*, [2]*t.heck@dipf.de*, [3]*c.schindler@dipf.de*
DIPF | Leibniz Institute for Research and Information in Education; Rostocker Straße 6, 60323 Frankfurt am Main (Germany)

Introduction

Citations are an important element of scientific communication, as they transparently show relationships between scientific publications, research data and their authors within the scientific community. Citation data is used in bibliometric and scientometric studies as evidence of internal scientific communication for the self-reflection of a discipline, for the evaluation and control of research performance and for research management (van Raan, 2019; Ball, 2020). In the past, the collection, processing and provision of scientific citation data was in the hands of a few commercial providers, such as Clarivate (Web of Science) and Elsevier (Scopus). Their use is based on licenses, which results in two major problems: Firstly, the commercial citation databases are subject to a fee and are not openly accessible. Secondly, those citation databases do not cover all disciplines to the same extent. As a result, these citation databases are only suitable for searching for literature and evaluating research to a very limited extent. This applies above all to the social sciences and humanities, which include almost all disciplines doing research about education, such as educational research, psychology, economics, and sociology (Moed, 2005; Singleton et al., 2015). Studies also show that reference lists in those databases are missing or are insufficient (Martín-Martín et al., 2018; Visser, van Eck, Waltman, 2021; Chi, 2014). In summary, educational research lacks exhaustive and high-quality citation data to improve literature search and disciplinary bibliometric studies.

Current research projects and network activities aim to contribute to open and networked citation data in science (Backes et al., 2024). Two examples of such approaches are the Initiative for Open Citations (I4OC) and OpenAlex. Our project Open Citation Data for Educational Research (OFFZIB) aligns with those initiatives and aims to extract citation data from open access publications in educational research and make them available via the central national German Education Index (FIS Bildung) (Botte, 2017). This meets the need for a more optimized literature search in the form of a semantic research graph in the database (Hocker et al., 2019) and at the same time offers the possibility of more detailed citation analyses in educational research. To reach this goal, we need to adapt an extraction algorithm to best perform with educational literature data and to establish new workflows to maintain the provision of the extracted data when the project has ended. To develop this extraction algorithm, knowledge must first be gained about how German education researchers cite, specifically in-text citations (Burbules, 2014). The specific research question is: Which citation styles (including special cases) exist in German educational research and are and there sub-disciplinary and document type-based differences?

Method

To investigate this question, a dataset was developed that represents the educational science publication landscape in Germany. The sample considers the different sub-disciplines of German educational research as well as the document types (data collection) and is coded regarding generally valid citation styles (coding).

Data Collection

The dataset shall represent the educational research publication landscape in Germany and thus is based on publications in the largest

disciplinary national open access repository peDOCS (Schindler & Butz, 2023). We aim to analyse at least 1% of the database peDOCS (~ 25,000 documents), thus determining a dataset of 400 documents. In the dataset, the ratio between the sub-disciplines (e.g. developmental psychology, educational sociology) and the existing three document types articles, books and collections (e.g. proceedings) are balanced according to the overall ratio in peDOCS. In addition, it was considered to ensure that the ratio of older and more recent publications as well as German and English documents in the peDOCS database is reflected.

Coding

The citation practices applied in the 400 documents are coded and analysed regarding common and standardised citation styles (e.g. APA citation style), but above all also with regard to styles specific for educational research. For example, special cases that cannot be assigned to a standardised citation style are citations of legal texts, which are then coded as an individual style. The documentation of the styles will be provided in an interoperable format to enable others to compare and reuse the collection for their own citation extraction.

Discussion

The citation practices of educational research are presented, compared and discussed against the background of other disciplines. Similarities and differences are highlighted. The result of the analysis is a comprehensive presentation of citation styles in educational research in Germany and their special formats. Furthermore, the results are discussed regarding challenges for citation extraction.

Outlook

Building on the results, the OFFZIB project will train the OUTCITE algorithm (Hosseini et al., 2019; Backes et al., 2024) to extract citations from educational open access publications. To make an active contribution to the development of a transdisciplinary and transnational citation inventory beyond the specific subject communities of educational research, the citation data will be given to the Open Citations Initiative. Therefore, a maintainable workflow will be established, which will also consider the workflows of the 30 partner institutes, which index and provide the literature for the German Education Index.

References

Backes, T., Iurshina, A., Shahid, M. A., & Mayr, P. (2024). Comparing Free Reference Extraction Pipelines. *International Journal on Digital Libraries*, 25(4), pp. 841–853. https://doi.org/10.1007/s00799-024-00404-6

Ball, R. (Hrsg.). (2020). *Handbook Bibliometrics*. De Gruyter Saur. https://doi.org/10.1515/9783110646610

Botte, A. (2017). 25 Jahre Fachinformationssystem (FIS) Bildung – eine einzigartige Kooperation. *Bibliotheksdienst* 51(8), pp. 651–663. https://doi.org/10.1515/bd-2017-0071

Burbules, N.C. (2014). The Paradigmatic Differences Between Name/Date and Footnote Styles of Citation. In: P. Smeyers, M. Depaepe (Eds.), *Educational Research: Material Culture and Its Representation*. Educational Research, vol 8. Springer, Cham. https://doi.org/10.1007/978-3-319-03083-8_13

Chi, P.-S. (2014). Which role do non-source items play in the social sciences? A case study in political science in Germany. *Scientometrics*, 101(2), pp. 1195–1213. https://doi.org/10.1007/s11192-014-1433-1

Hocker, J.; Veja, C., Schindler, C.; Rittberger, M., (2019). Establishing semantic research graphs in humanities' research practice. In: C. Draude, M. Lange & B. Sick (Eds.), *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft* (Workshop-Beiträge). Bonn: Gesellschaft für Informatik e.V. (pp. 169-174). https://doi.org/10.18420/inf2019_ws18

Hosseini, A., Ghavimi, B., Boukhers, Z., & Mayr, P. (2019). EXCITE - A toolchain to extract, match and publish open literature references. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2019*, pp. 432–433. https://doi.org/10.1109/JCDL.2019.00105

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), pp. 1160-1177. https://doi.org/10.1016/j.joi.2018.09.002

Moed, H. F. (2005). *Citation analysis in research evaluation*. Information Science and Knowledge Management: Bd. 9. Springer. https://doi.org/10.1007/1-4020-3714-7

Schindler, C. & Butz, A. (2023). peDOCS - ein Fachrepositorium in der Bildungsforschung mit Kooperationsnetzwerk für Open Access. In H. Ertl & B. Rödel (Eds.), *Offene Zusammenhänge: Open Access in der Berufsbildungsforschung* (Berichte zur beruflichen Bildung, pp. 236-242). Bonn: Bundesinstitut für Berufsbildung. URL: https://www.bibb.de/dienst/publikationen/de/18249

Singleton, K., Kuhberg-Lasson, V., Sondergeld, U., & Schultheiß, J. (2015). Publikationen der Bildungsforschung. In A. Botte, U. Sondergeld, & M. Rittberger (Eds.), *Monitoring Bildungsforschung: Befunde aus dem Forschungsprojekt "Entwicklung und Veränderungsdynamik eines heterogenen sozialwissenschaftlichen Feldes am Beispiel der Bildungsforschung"* (pp. 69–106). Klinkhardt.

Van Raan, A. (2019). Measuring Science: Basic Principles and Application of Advanced Bibliometrics. In: W. Glänzel, H.F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Switzerland: Springer.

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1), pp. 20–41. https://doi.org/10.1162/qss_a_00112

https://doi.org/10.51408/issi2025_180

Portuguese Scientific Production: Volume indicators

Catarina Carreira[1], Cristiana Agapito[2]

[1]*catarina.carreira@dgeec.medu.pt*, [2]*cristiana.agapito@dgeec.medu.pt*
Directorate General of Education and Science Statistics (DGEEC), Av. 24 de Julho, n.º 134, 1399-054 Lisbon (Portugal)

Introduction

In Portugal, one of the main goals of data related to scientific production is to provide essential tools for the processes of diagnosis, evaluation, and monitoring of national scientific system for the implementation of public policies. Bibliometric analysis uses data on the number of publications and citations, allowing us to understand the dynamics of different research areas, as well as observe the research outcomes of institutions, researchers, and countries within scientific system. These data also allow the identification of national and international collaboration networks and the flow of knowledge among them, thereby providing a perspective on the globalization of science. DGEEC presents some indicators related to scientific publication volume, namely the number of publications in which at least one of the authors is affiliated with a national institution. The data include all types of documents, except for the indicators on international comparisons and Sustainable Development Goals (SDGs), where the bibliographic records was limited to citable documents classified as articles and reviews. The presented data are the result of analysis conducted on the international platform *InCites*, a product of Clarivate Analytics, and are based on the *Web of Science* (core collection) data source.

A global scenario

In 2023, the number of publications with Portuguese affiliation indexed in the Web of Science was 27,646, an increase of 33% compared to 2013. Of these publications, 66% were open access, demonstrating the consolidation of new publishing practices within the scientific community, in which research results are made available free of charge and online. Considering publications classified by scientific area (FORD), there has always been a predominance of the natural sciences, medical and health sciences and engineering and technology sciences, something that should not be dissociated from the fact that there is a greater representation of these areas in this data source.

The highest average annual growth rates between 2013 and 2023 were 4.5% for publications classified as agricultural and veterinary sciences, 4.3% for publications in the social sciences and 4.0% for publications in medical and health sciences.

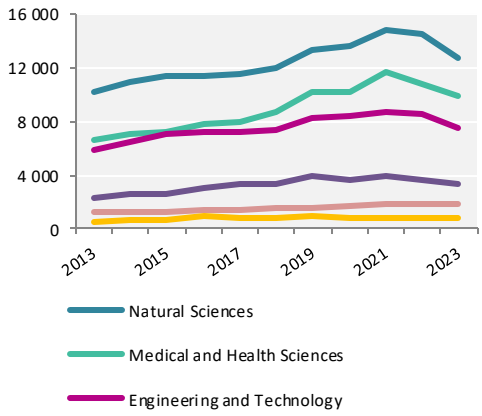


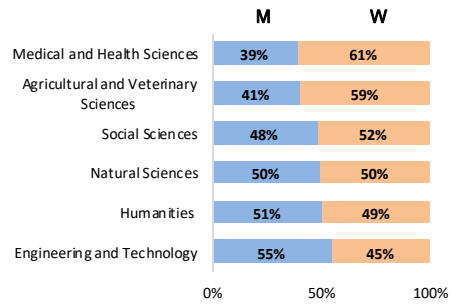
Figure 1. Portuguese Publications, by scientific area (2013-2023).

In 2023, more than half (57%) of Portuguese publications were co-authored with institutions from other countries, an increase of 12 percentage points compared to 2013 (45%). In 2023, Portugal collaborated with 198 countries, with Spain, the United Kingdom and the United States of America having the highest number of co-authored publications.

Gender Indicators

Gender statistics and indicators are important tools to promote gender equality and measure gender gaps (EIGE, 2019). Between 2018 and 2022, authors with Portuguese affiliation who published on the Web of Science were mostly women (53%). Men mostly play the role of last author (55%), correspondence author (51%) and unique author (58%). In the articles published, women assume a greater weight in the position of 1st author (53%).

Publications in Engineering Sciences and Technology have mostly male authors (55%). The Medical Sciences and Health Sciences have the highest % of women as authors (61%).



* In publications from 2018 to 2022, 50% of authors with at least one Portuguese affiliation were classified by sex. The data presented are based on these classified authors

Figure 2. Portuguese affiliated author, by sex and scientific area (2018-2022).

Sustainable Development Goals (SDGs)

The Sustainable Development Goals (SDGs) were established by the United Nations (UN) in 2015, with the aim of guiding actions towards a more sustainable future for people and the planet. Adopted by all Member States of the United Nations, the 2030 Agenda represents a call to action by all countries - developed and developing - for a global partnership around common goals and targets to end poverty, protect the planet, and ensure peace and prosperity by 2030.

Given the fundamental role of science in achieving the goals and targets associated with the SDGs, it is important to publish data on Portuguese scientific production in this area. The data presented refers to the volume of Portuguese scientific production, indexed in the Web of Science, published between 2019 and 2023, referring to each of the SDGs. 96% of Portuguese publications belong to Objective 3 – Good Health and Well-Being. Secondly, 16% of the publications are related to Objective 14 – Life Below Water and Objective 15 – Life on Land.

International comparison

To compare the volume of scientific production between European Union countries, a normalised indicator is presented. The normalization considers the population aged between 25 and 64 with tertiary education.

This indicator seeks to measure the intensity of the production of scientific articles in the country among the population group that, from the outset, would be most able to participate directly in scientific activities, in other words, it is an indicator of the intensity of scientific production among the "target" population group.

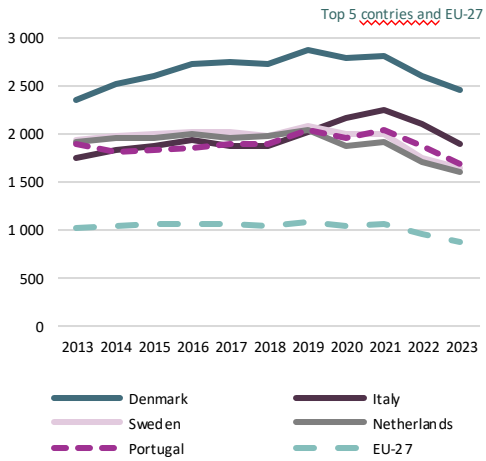


Figure 3. Number of publications per 100,000 inhabitants, aged between 25 and 64, with Tertiary Education.

In this indicator, Portugal is in 5th place with an average of 1.882 publications per 100.000 inhabitants aged between 25 and 64 with Tertiary Education. In the top two places are Denmark and Italy, with an average of 2.645 and 1.953 publications respectively.

This short summary of the Portuguese panorama regarding the volume of scientific publications is just a sample of indicators that have been made available in more depth in general and thematic publications that can be

consulted on the DGEEC website at https://www.dgeec.medu.pt/.

References

Direcção-Geral de Estatísticas da Educação e Ciência. *Portuguese scientific production (several publications)* Lisbon: DGEEC.

Elsevier. (2021). *Gender in the Portugal research arena: A case study in European leadership*.

European Commission. (2015). *Analysis of bibliometric indicators for European policies (2000-2023)*. Science-Metrix.

European Institute for Gender Equality. (2019). *Gender statistics and indicators*. Publications Office of the European Union.

OECD & SCImago Research Group (CSIC). (2016). *Compendium of bibliometric science indicators*. OECD. Retrieved from http://oe.cd/scientometrics

Vieira, A., & Fiolhais, C. (2015). *Science and technology in Portugal: Metrics and impact (1995-2011)*. Francisco Manuel dos Santos Foundation.

Vieira, E. S., & Gomes, J. A. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, *4*(1), 1–13. https://doi.org/10.1016/j.joi.2009.06.002

Vieira, E., Mesquita, J., Silva, J., et al. (n.d.). *The evolution of science in Portugal (1987-2016)*. Francisco Manuel dos Santos Foundation.

Waltman, L., & Noyons, E. (2018). *Bibliometrics for research management and research evaluation: A brief introduction*. CWTS – Leiden University.

https://doi.org/10.51408/issi2025_165

Productivity and Impact Patterns in Scientific Careers

Kaile Gong

gongkaile@njnu.edu.cn
School of Journalism and Communication, Nanjing Normal University, Nanjing, Jiangsu 210097 (China)

Introduction

The patterns of scientific careers have long been of interest to scientometrics (Sinatra et al., 2016). Mobility, especially international mobility, is widely recognized to have a significant impact on the development of scientific careers (Netz, Hampel & Aman, 2020). Although existing research has made many beneficial discoveries, they often rely on some small samples of elite scientists (e.g. Nobel Prize winners), with insufficient exploration of broader patterns and the role of international mobility in them. Therefore, this study takes PubMed as the data source, adopts the method of time series clustering to reveal multiple patterns of productivity and impact in the academic careers of 67,201 scientists, and then uses the Chi-square test to analyze the influence of international mobility.

Methodology

Data

The data was collected from the PubMed Knowledge Graph 2.0 (PKG 2.0), which is an open dataset built by Xu et al. (2024). PKG 2.0 provides PubMed-indexed papers published before 2024 and has high-quality author disambiguation with an F1 score of 96.24%. More importantly, it integrates multi-source data and maps partial PubMed authors to Orcid scholars, which offers accurate information about scientists' education and employment. Thus, both publications and international mobility can be identified and analyzed based on PKG 2.0. Since PubMed is a biomedical and life science database, this study's findings are applicable to this field.

In order to ensure that the selected scientists have a sufficiently long and continuous career and that their mobility can be identified via Orcid, this study draws on the approach of Sinatra et al. (2016) and applies four inclusion criteria: (1) the scientists should have at least a 30-year publication career; (2) the scientists should author at least one paper every 5 years; (3) the scientists should publish at least 30 papers; and (4) the scientists should have education and employment records in Orcid. Finally, the samples for analysis include 67,201 scientists and 8,769,452 papers they published from 1936 to 2023.

Productivity, impact and international mobility

Productivity refers to the number of papers published within a certain time range, so for each scientist, the number of papers published each year is counted to obtain the yearly publication sequence.

Impact refers to the citation impact of the most cited paper published by an author within a time range, specifically, it's defined as the highest 5-year citation count of a paper published within that time. Thus, for each scientist, the 5-year citation count of all papers published before 2018 is first counted by considering the 5-year citation window, then the most cited paper in each year is found and its 5-year citation count is used as the impact indicator in that year. Finally, the yearly impact sequence of each scientist is obtained.

International mobility is identified for each scientist based on the presence of two or more different countries in their Orcid education and employment records.

Time series clustering

Dynamic time warping (DTW) and K-medoids are combined as the time series clustering method to detect productivity and impact patterns in scientists' careers. DTW is the most popular and widely accepted method for measuring the similarity between time series data with different lengths (Ao et al., 2023). K-Medoids is a clustering algorithm

similar to K-means, but it selects real points existing in the dataset as cluster centroids instead of calculating the average of all points, which makes K-Medoids more robust in handling noise and outliers (Arora & Varshney, 2016).

Let's take productivity patterns clustering as an example to briefly explain the clustering process: firstly, the Python package TAIDistance (Meert et al., 2022) is used to calculate the pairwise DTW distance between the yearly publication sequence of all scientists to form the distance matrix, and then the distance matrix is input into the K-medoids to implement clustering. The elbow method based on inertia value is used to determine the number of clusters, which is 3 in the clustering of productivity patterns and 4 in the clustering of impact patterns.

The influence of international mobility

The Chi-square test is used to analyze whether there are significant differences in the distribution of mobile and non-mobile scientists in different productivity (impact) patterns.

Results

The productivity and impact patterns in scientific careers are respectively shown in Fig. 1 and Fig. 2. The green lines in the figures are the sequences of K-Medoids centroids and the red lines are the modified centroids using the DTW Barycenter Averaging (DBA) algorithm, which can better represent each pattern.

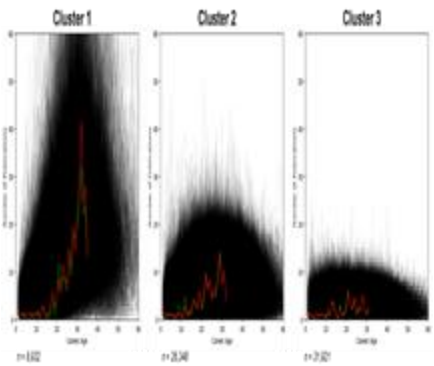


Figure 1. Productivity patterns in.

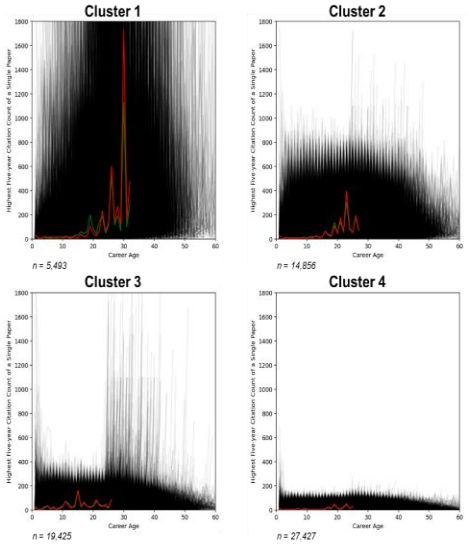


Figure 2. Impact patterns in scientific careers.

Table 1. Influence of mobility on productivity patterns.

		Productivity patterns			χ^2	p
		Cluster 1	Cluster 2	Cluster 3		
Mobility types	Mobility	6,311	16,240	16,659	1172.362	0.000***
	%Mobility	16.1%	41.4%	42.5%		
	%Cluster *i*	70.7%	61.6%	52.2%		
	Non-mobility	2,621	10,108	15,262		
	%Non-mobility	9.4%	36.1%	54.5%		
	%Cluster *i*	29.3%	38.4%	47.8%		

*** $p<0.001$

Table 2. Influence of mobility on impact patterns.

		Impact patterns				χ^2	p
		Cluster 1	Cluster 2	Cluster 3	Cluster 4		
Mobility types	Mobility	3,809	9,869	11,536	13,996	1285.608	0.000***
	%Mobility	9.7%	25.2%	29.4%	35.7%		
	%Cluster *i*	69.3%	66.4%	59.4%	51.0%		
	Non-mobility	1,684	4,987	7,889	13,431		
	%No-mobility	6.0%	17.8%	28.2%	48.0%		
	%Cluster *i*	30.7%	33.6%	40.6%	49.0%		

*** $p<0.001$

It can be concluded from Fig. 1 that the productivity patterns include three types: high peak (Cluster 1), moderate peak (Cluster 2), and low fluctuation (Cluster 3), and peaks often appear in the late stage of scientific careers. According to Fig. 2, the impact patterns include four types: high peak (Cluster

1), moderate peak (Cluster 2), low peak (Cluster 3), and flat (Cluster 4), and the age of the peak is advanced with the peak value decreasing. In addition, it is found that no matter productivity or impact, there's little difference between various patterns during the first 1/3 career, but after that, there's a clear divergence. The potential policy implication of the findings is that we need to be more patient with scientific career development, and what triggers the divergence deserves future attention.

Table 1 and Table 2 show that mobile and non-mobile scientists have significant differences in the distribution of productivity and impact patterns, and mobile scientists are more likely to achieve relatively higher peaks, which can be inferred that mobility is beneficial to scientific career development.

Acknowledgments

This study was funded by the National Social Science Fund of China (No. 23CTQ032).

References

Ao, W., Lyu, D., Ruan, X., Li, J., & Cheng, Y. (2023). Scientific creativity patterns in scholars' academic careers: Evidence from PubMed. *Journal of Informetrics, 17*(4), 101463.

Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78, 507-512.

Meert, W., Hendrickx K., Van Craenendonck, T., Robberechts, P., Blockeel, H., & Davis, J. (2022). DTAIDistance (Version v2). *Zenodo*, http://doi.org/10.5281/zenodo.5901139.

Netz, N., Hampel, S., & Aman, V. (2020). What effects does international mobility have on scientists' careers? A systematic review. *Research Evaluation, 29*(3), 327-351.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A. L. (2016). Quantifying the evolution of individual scientific impact. *Science, 354*(6312), aaf5239.

Xu, J., Yu, C., Xu, J., Ding, Y., Torvik, V. I., Kang, J., ... & Song, M. (2024). PubMed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *arXiv preprint arXiv:2410.07969*.

https://doi.org/10.51408/issi2025_187

Publications at the Intersection of Academia and Market: Unpacking Scholarly Outputs of University-Industry Collaboration in Brazil

Gabriel Falcini[1], Sergio Luiz Monteiro Salles Filho[2], Yohanna Juk[3]

[1]*falcini.gabriel@gmail.com*, [2]*sallesfi@unicamp.br*, [3]*yohannajuk91@gmail.com*
Science and Technology Policy Department, University of Campinas, R. Carlos Gomes, 250 - Cidade Universitária, Campinas - SP, 13083-855, CEP 13083-855, Campinas (Brazil)

Introduction

University-Industry Collaboration (UIC) offers significant benefits to stakeholders, stimulates market competitiveness, and strengthens the economy. Promoting this collaboration type is a strategy that has captured the attention of policymakers in the field of Science, Technology, and Innovation (STI) worldwide.

The motivations for the agents involved in a partnership between Research Organisations (ROs) and companies are diverse. For the industry, incentives may arise through the reduction of research and development (R&D) costs, access to the latest advances in research knowledge, utilisation of cutting-edge laboratory infrastructure, enhancement of internal skills, and increased competitiveness (Hagedoorn, Link, & Vonortas, 2000; Kroll, 2016). On the other hand, for ROs, motivations include access to industrial production infrastructure, opening venues for technology and knowledge transfer, sharing the risk of projects and opening of new research streams or the deepening of existing ones (Salles-Filho et al., 2021). Moreover, for the economy and society, UIC can unfold benefits such as increased national public and private investment in R&D, exchange and sharing of scientific and technological knowledge among social agents, and the acceleration of innovative and technological solutions that can enhance the quality of life and international competitiveness (Hagedoorn, Link, & Vonortas, 2000).

In the face of the abovementioned benefits, it is necessary to understand the barriers that hinder the realisation of University-Industry Collaboration, especially in the Brazilian context. According to the 2023 Global Innovation Index (WIPO, 2023), Brazil ranks 78th out of 132 countries in the global UIC ranking. Moreover, this is not a recent scenario but a historical difficulty that has persisted in stagnation, as asserted by Faria (2021) using data from the Brazilian Industrial Research of Technological Innovation (PINTEC) until 2017. According to Salles-Filho et al. (2021), the factors obstructing collaboration can be of two natures: asymmetries in strategic orientations and incentives; and transaction barriers regarding intellectual property and structural bureaucracy.

Well-founded public policies have the potential to mitigate existing barriers and provide incentives for collaboration. The Organisation for Economic Co-operation and Development (OECD, 2002) outlines a series of public policies that can be established to encourage UIC: financial incentives for collaborative research, cooperative research centres, public seed capital funds, publicly funded intermediaries, among others. In this sense, our research aims to contribute to the understanding of the UIC landscape in Brazil, a topic that receives little attention in STI research in the country. Through bibliometric indicators, we seek to better understand the barriers to collaboration, particularly for ROs.

Research Design

To investigate the Brazilian University-Industry Collaboration scenario for technological innovation, we will use bibliometric indicators from OpenAlex, covering articles published from 2012 to 2022. This period was selected to ensure adequate maturation of scholarly output indicators, such as citation and readership

counts. The study sample will consist of publications with at least one company and one university among the authors' affiliations, with all involved organisations being Brazilian.

For this set of publications, we will analyse the following indicators: number of authors per publication, number of collaborating institutions, number of funders, number of cited references, and open access availability. Additionally, we will investigate the scientific impact of the publications based on the impact factor of the journals, the number of citations received, citations in patents, and citations in policy documents (figures extracted from Overton). Additional engagement and visibility metrics will be analysed, including readership, views or downloads, and mentions on social media, as recorded by Altmetric.

To contextualise the results, we will compare these indicators with those of a matching control group derived from the sample: publications with Brazilian affiliations in the same thematic areas and period but without industry collaboration.

Preliminary results and expected contributions

We conducted an initial search in OpenAlex using the following criteria: articles published between 2012 and 2022 that include at least one institution classified as a "company" and at least one classified as "education". Furthermore, to avoid bias in the scientific impact indicators, all participating institutions had to be based in Brazil. Other types of publications, such as reviews, book chapters, and dissertations, were excluded to ensure greater consistency in the analysis. As a result, our preliminary sample comprises 3,565 articles, jointly published by 161 companies ("company", in OpenAlex) and 327 universities ("education", in OpenAlex). Other collaborating institutions also appear in the sample, such as "nonprofit", "government", and "healthcare" — all of which are Brazilian.

OpenAlex organises "concepts" into five hierarchical levels, where the lower levels represent broader areas of knowledge, while the higher levels correspond to more specific topics. For example, level 0 may include the concept "Biology", whereas level 3 may include "Plant Pathology". These concepts are assigned to publications based on a score ranging from 0 to 1. The higher the score, the greater the relevance of that concept to the content of the publication. From the preliminary sample, we extracted all level 3 concepts associated with the articles, provided they had a score above 0.7. We then searched the database for all articles linked to the same concepts—also with a score greater than 0.7—which are affiliated exclusively with Brazilian institutions and which, importantly, never contain both an institution of type "education" and an institution of type "company" simultaneously. This process yielded a preliminary comparison group consisting of 69,800 articles that are highly similar to those in the sample (sharing the same highly relevant concepts), but which were not published in University–Industry Collaboration.

By comparing the scientific impact indicators of the sample and the comparison group, we expect to find a lower scientific impact for publications derived from UIC, as these tend to focus more on the internal needs of companies and less on broader scientific questions, as corroborated by the literature (Ankrah & AL-Tabbaa, 2015; Pujotomo et al., 2023; Hong & Su, 2013). On the other hand, we may find a greater altmetric impact for the sample.

Bringing this evidence to the Brazilian context is important for understanding the motivations and barriers, particularly for universities, enabling policymakers to make more informed decisions when designing policies to foster UIC. We also hope to contribute to the understanding of UIC in the Global South, given that Brazil shares similarities with other developing countries in terms of technological innovation in industry.

Limitations and future studies

A potential limitation of the study lies in the inconsistencies observed in OpenAlex, which are more frequent when compared to databases such as WoS and Scopus, given its less stringent editorial curation. Nevertheless, OpenAlex was selected due to being an open and freely accessible database, as well as for its broader coverage of countries from the Global South, including Brazil. Random checks were conducted on the preliminary data, which showed consistency in the

classification of both institution types and concepts. However, we do not rule out the possibility that a change of database may become necessary, should such inconsistencies prove to be significant.

Further studies are needed to deepen the analysis, for instance, by examining patent indicators to assess the technological impact of these collaborations, as well as conducting a comparative study on the scientific impact of UIC in Brazil, contrasting it with other developing countries similar to Brazil and with developed countries, which typically exhibit higher levels of industrial innovation. Similar studies conducted using the WoS and Scopus databases may also offer valuable insights for comparative purposes.

Acknowledgments

We acknowledge the São Paulo Research Foundation (FAPESP) for its support through a doctoral scholarship (grant 2021/11476-2) and a postdoctoral fellowship (grant 2021/06285-3).

References

Ankrah, S. & Al-Tabbaa, O. (2015). Universities–industry collaboration: A systematic review. *Scandinavian Journal of Management*, v. 31, n. 3, p. 387–408.

Faria, P. (2021). Cooperação Universidade - Indústria No Brasil: suas características e desafios, a partir da PINTEC (2000 - 2017). *Revista Multiface Online* (pp. 5–36). v. 9, n. 1.

Hagedoorn, J., Link, A., Vonortas, N. (2000). Research partnerships. *Research Policy* (pp. 567–586). v. 29, n. 4–5.

Hong, W., Su, Y. (2013). The effect of institutional proximity in non-local university–industry collaborations: An analysis based on Chinese patent data. *Research Policy*, v. 42, n. 2, p. 454–464.

Kroll, H. (2016). Supporting New Strategic Models of science-industry R&D collaboration: A review of global experiences (p. 44).

OECD. (2002). Benchmarking Industry-Science Relationships. OECD.

Pujotomo, D., et al. (2023). University–industry collaboration in the technology development and technology commercialization stage: a systematic literature review. *Journal of Applied Research in Higher Education*, v. 15, n. 5, p. 1276–1306.

Salles-Filho, S., et al. (2021). Effectiveness by Design: Overcoming Orientation and Transaction Related Barriers in Research-Industry Linkages. *Revista de Administração Contemporânea* (p. 22). v. 25, n. 5.

WIPO. (2023). Global Innovation Index 2023. Geneva, Switzerland: World Intellectual Property Organization (WIPO).

https://doi.org/10.51408/issi2025_190

Quality Evaluation of Scientific Journals in the Open Science Context

Lei Li[1], Yue Hu[2], Hui Peng[3], Shi Chen[4]

[1]leili@bnu.edu.cn, [2]202221260043@mail.bnu.edu.cn, [3]hui_peng@mail.bnu.edu.cn, [4]kjchensh@bnu.edu.cn
School of government, Beijing Normal University, Beijing (China)

Introduction

Open science is a global initiative aimed at enhancing the quality, transparency, and societal impact of scientific research. It seeks to foster reproducibility, informed policymaking, and public trust in science (UNESCO, 2021). As the open science movement grows, academic journals—key platforms for disseminating scholarly knowledge—must adapt by aligning their operations with open principles. However, this shift has introduced concerns about journal quality, especially regarding peer review rigor, ethical standards, and the potential prioritization of commercial interests over scientific integrity.

Scientific and technical journals, which frequently lead in open access adoption due to the nature of their content, bear particular responsibility. They reflect a country's scientific capacity, contribute to international competitiveness, and influence the direction of research and policy. If these journals compromise on quality, the consequences can be severe, including misleading scholars and decision-makers and eroding public trust in scientific communication. As such, evaluating and improving journal quality in the context of open science is both urgent and essential.

Traditionally, journal evaluation has relied on citation metrics or alternative bibliometric indicators. While these are useful for measuring scholarly impact, they provide only a partial view and are often outcome focused. They fail to capture the entire publishing lifecycle and overlook key elements such as openness, transparency, service quality, and ethical practices. Therefore, a more comprehensive and process-oriented evaluation system is necessary.

This study aims to construct a multidimensional framework for assessing the quality of scientific journals under the open science paradigm. It considers the full lifecycle of scholarly publication—from manuscript submission and peer review to dissemination and societal influence—allowing for a more nuanced understanding of journal performance.

Construction of the Evaluation Framework

To align with the open science agenda, this study began by reviewing definitions, policies, and practices from academic literature and major publishers (Vicente-Saez & Martinez-Fuentes, 2018; Elsevier, 2025; Saha et al., 2003). Based on this foundation, a preliminary set of evaluation indicators was drafted. The Delphi method was used to solicit feedback from 30 experts with experience in open science, leading to a refined indicator set through two rounds of expert consultation. The framework is rooted in the principles of Total Quality Management (TQM), dividing journal quality into two overarching dimensions: product and service.

The product dimension assesses the openness and integrity of research outputs published by the journal. The service dimension evaluates the journal's efforts to support authors, readers, and the broader public through open science practices and knowledge dissemination.

Expert Evaluation and Weighting

Using the Delphi method, the indicator system was refined through two rounds of consultation with 30 experts. In round one, 12 valid responses were received. While all indicators were retained, experts suggested clearer definitions and broader coverage. In round two, 7 experts affirmed the improvements. Indicators were then assigned

weights using the Analytic Hierarchy Process (AHP), based on the averaged importance scores. A consistency check ensured the validity of the final weighting scheme. The final indicators and their weights are shown in Table 1.

Empirical Analysis: Open-Access Journals in Optics

To validate the framework, an empirical study was conducted using open-access journals in the field of optics. Journals were retrieved from the DOAJ database using the keyword "Light" and cross-referenced with the 2022 Journal Citation Reports (JCR). Of 28 initially identified journals, 19 met the inclusion criteria (available website and JCR index).

Each journal was assessed according to the framework. Binary scoring (1 = present; 0 = not present) was applied for qualitative indicators based on website information. Academic impact was measured using normalized impact factors. Social impact was derived from Altimetric scores calculated for papers published between 2020–2022. Publication transparency and other services were evaluated based on publicly available editorial and operational information.

Results

The top three journals—*Optica*, *Optics Continuum*, *and Optical Materials Express*—are all published by the Optica Publishing Group. These journals consistently support open peer review, require data availability statements, and promote publications through comprehensive outreach. They exemplify strong alignment with open science principles across both product and service dimensions.

Mid-tier journals, including *EPJ Quantum Technology and Photoacoustics*, performed reasonably well but lacked features like publication bias statements or robust open review processes.

Table 1. Evaluation Indicator Framework for Scientific Journals in the Open Science Context.

Dimension	Indicator and Weight (%)	Explanation
Product	Open Research Process (10.83)	Supports pre-registration of research and ensures transparency in the entire research process, from the start of the project to its completion. This includes research work, implementation plans, technical routes, analytical methods, experimental processes, and public engagement.
	Preprint Licensing (19.61)	Allows authors to publicly share manuscript drafts on designated preprint platforms before formal publication, based on well-established preprint copyright, licensing, ethics, privacy, and general guidelines.
	Open Peer Review (15.06)	Disclosure of reviewers' identities, public review comments, and the opportunity for broader community input in evaluations.
	Open Scientific Outputs Related to Publications (8.10)	Includes raw research data, software, source code, materials, hardware designs, protocol workflows, images, charts, multimedia materials, and other related scientific outputs.
	Open Repository (7.31)	A platform that offers access to relevant materials (e.g., research data, scientific outputs) in formats that are user-friendly, machine-readable, and interoperable with open research infrastructures.
	Paper Content Quality (2.50)	Strict checks for academic misconduct, ensuring that all published papers adhere to strict data citation rules and quality standards.
	Academic Impact (2.88)	The use of papers, including views, downloads, and citations of abstracts and full text.
	Social Impact (8.71)	The number of shares, retweets, likes, and other forms of engagement on new media, along with political and economic impacts.
Service	Author Open Policy Service (3.96)	Describes the journal's open policies in the submission guidelines and provides a checklist of submission requirements under these policies, including explanations for special cases.
	No Publication Bias Statement (4.58)	The journal declares that the significance and novelty of research results are not the sole criteria for publication. During the review process, the journal does not consider the outcomes of the research. It accepts replication studies and registered reports of innovative research, treating these as regular submission options.
	Publication Transparency (1.96)	Provides detailed information about the process from submission to peer review to final publication, including initial decision times, average review times, number of reviews, and geographical distribution of editors and reviewers.
	Diverse Publication Formats (2.50)	A variety of publishing formats, such as XML/HTML web publishing, multimedia publishing, semantic publishing, enhanced publishing, etc.

	Diverse Promotion Services (3.12)	Comprehensive use of various promotional methods, such as targeted email campaigns, promotion via different new media platforms, and hosting public academic conferences and outreach activities to promote academic exchange and collaboration.
	Online Communication Platform (2.62)	Provides online platforms or social media for the public to discuss research processes, data, methods, and publications.
	Open Science Outreach Activities (3.07)	Collaborates with universities and research institutions to offer lectures or training sessions that explain open peer review, open publishing, and other related topics to improve the utilization of open academic resources.
	Open Resource Usage Instructions (3.20)	Provides readers with detailed explanations of the open resources available, including guidelines for using the resources, ensuring accessibility and ease of use.

Lower-ranked journals such as *Light: Science & Applications* demonstrated limited engagement in key areas like open research processes and community outreach, despite offering open access.

Common weaknesses across all journals included insufficient support for pre-registration, limited use of multimedia formats, and the general lack of open peer review practices. These gaps suggest a need for broader adoption of open science infrastructure and cultural changes in publishing norms.

Conclusion and Future Work

This study presents a comprehensive, empirically tested framework for evaluating journal quality under the open science paradigm. It integrates both outcome-based and process-based metrics and accounts for the full lifecycle of research dissemination. The results underscore the importance of transparency, data sharing, and community engagement as essential elements of journal quality in the digital age. By embracing a multidimensional evaluation perspective, journals can better align with the principles of open science, thereby fostering a more transparent, equitable, and impactful scholarly communication ecosystem.

In subsequent research, the indicator framework will be further improved, its empirical scope broadened through evaluations of scientific and technological journals across various fields, the definitions and applications of each indicator will be continually refined and specified, and by comparing it with existing evaluation models, the credibility and generalizability of the indicator framework will be enhanced. Additionally, comparative empirical analyses of journals from different countries could be conducted, drawing on best practices to promote the development of high-quality open science journals.

Acknowledgments

This work was supported by Beijing Natural Science Fund Project (No. 9232013).

References

UNESCO. (2021). *Recommendation on Open Science*, Retrieved January 16, 2025 from: https://www.unesco.org/en/open-science/about.

Vicente-Saez, R. and Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of business research*, *88*, 428-436.

Saha, S., Saint, S. and Christakis, D.A. (2003). Impact factor: a valid measure of journal quality?. *Journal of the Medical Library Association*, *91*(1), 42-46.

Elsevier. (2025). *Advancing open access to knowledge*, Retrieved January 16, 2025 from: https://www.elsevier.com/open-access

https://doi.org/10.51408/issi2025_178

Recent Advance of Text Mining in LIS: A bibliometric review

Siqi Hong[1], Guo Chen[2]

[1] 3402202918@qq.com, [2] delphi1987@qq.com
Nanjing University of Science and Technology, No. 200 Xiao Ling Wei, Nanjing, Jiangsu (China)

Introduction

Text mining has become an essential tool in Library and Information Science (LIS), yet systematic reviews remain scarce. Early reviews mainly provided technical overviews, while recent research has expanded into specific application areas.

Based on this gap, this study analyzes text mining research in LIS from 2022 to 2024. We first apply topic modeling to identify key research directions, then focus on three core questions:

1) What types of texts are studied?
2) What technologies are used?
3) What are the main application scenarios?

To answer these, we propose a "Text–Technology–Scenario" three-dimensional framework that examines LIS text mining from the perspectives of research objects (what), methodologies (how), and application value (why), offering a structured view of its current landscape and future trends.

Data and methods

This study analyzed 1,726 text mining-related papers published between 2022 and 2024 in 39 authoritative LIS journals (23 Chinese, 16 English). A Python-based tool was used to collect bibliographic data, followed by keyword screening and manual verification. Chinese papers were automatically translated using a LLM API.

To identify research trends, we applied BERTopic and resulting in 17 major topics (Figure 1).

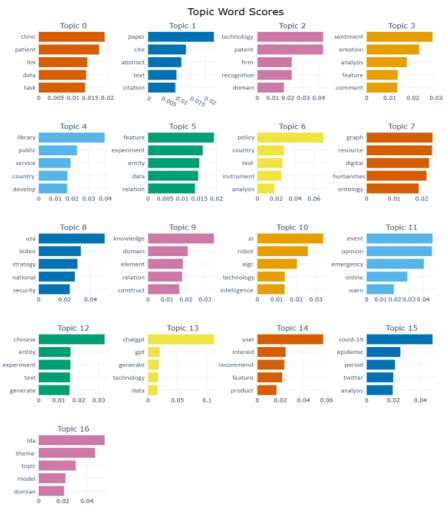


Figure 1. Visualization of topic word distribution.

We then conducted content analysis to interpret each topic. Based on this, we developed a three-dimensional framework (Figure 2) to categorize the findings into three analytical perspectives.

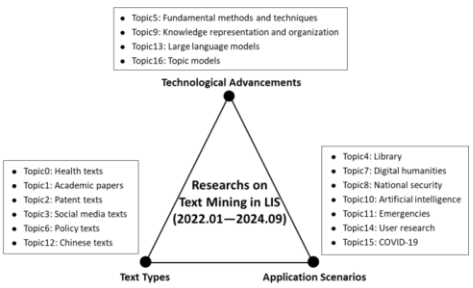


Figure 2. A Three-Dimensional Framework of Text Mining Research in LIS.

Results and discussion

Text Type Perspective

Different types of texts present distinct challenges and research priorities in LIS text mining. Table 1 summarizes major topics associated with six representative text types.

Table 1. Summary of Topics by Text Type.

Topic	Text Type	Key Focus
T0	Health Texts	Clinical decision support, disease prediction, knowledge services
T1	Academic Papers	Word/citation/topic-level mining, scientific evaluation
T2	Patent Texts	Tech evolution, opportunity detection, entity mapping
T3	Social Media	Sentiment analysis, misinformation detection
T6	Policy Texts	Cross-national analysis, coordination issues
T12	Chinese Texts	Classical/ancient texts, cultural NLP tasks

In health texts, Chinese research emphasizes knowledge services, while English studies focus on clinical applications—with LLMs widely applied in both. For academic papers, the focus has shifted from feature extraction to semantic understanding, with growing use of bibliometric methods for scientific evaluation. Patent research highlights cross-domain opportunities and merges patent with social media data for trend prediction. Social media research has moved from public opinion tracking to sentiment analysis and misinformation detection. Policy text mining is more active in Chinese but remains methodologically limited. Research on Chinese texts focuses on linguistic heritage such as classical Chinese and minority languages.

Technological Perspective

Recent LIS research has actively explored new methods to enhance semantic understanding and task performance. Table 2 highlights four representative topics from technological standpoint.

Table 2. Summary of Topics by Technological Focus.

Topic	Technology Focus	Key Themes
T5	Core Methods	Prompt learning, multimodal fusion, NER, classification
T9	Knowledge Representation	Entity/tuple/document-level representation for organization and application, interdisciplinary knowledge mining
T13	Large Language Models (LLMs)	ChatGPT applications, opinion mining
T16	Topic Modeling	LDA+BERT, trend analysis, user modeling

LIS text mining shows dual momentum: performance breakthroughs and knowledge-centered exploration. Core methods are enhanced via deep learning and multimodal fusion, especially in low-resource settings. Meanwhile, knowledge representation is advancing from carrier-level to semantic-level modeling.

LLMs like ChatGPT empower tasks such as summarization and entity extraction, while also raising concerns around ethics and hallucinations. Topic models continue evolving through integration with BERT and transfer learning, expanding to new domains like policy and culture. Overall, LIS research is shifting toward intelligent, multimodal, and domain-adaptive methods.

Application Scenario Perspective

LIS researchers are applying text mining to a wide range of practical domains with diverse goals and methods. Table 3 outlines seven prominent application areas identified in the literature.

Table 3. Summary of Topics by Application Scenario.

Topic	Application Scenario	Key Focus
T4	Libraries	Resource organization, service design
T7	Digital Humanities	Cultural heritage mining, ontology construction
T8	National Security	Policy analysis, strategic insight
T10	Artificial Intelligence	Chatbots, AIGC applications, ethical issues
T11	Emergencies	Opinion evolution, knowledge graphs for events
T14	User Research	Recommendation, demand mining, satisfaction analysis
T15	COVID-19	Sentiment evolution, multilingual health texts mining

In libraries, it supports the smart organization of digital resources; in digital humanities, it aids the analysis of cultural heritage; and in national security, it enhances policy and intelligence research. Emergency-related studies widely adopt ontologies and knowledge graphs to enable semantic understanding and causal reasoning, improving event inference and decision-making. User research has shifted from global to short-term interest modeling, with applications expanding beyond e-commerce to areas such as academic citation and community Q&A. In the context of AI, the rise of generative models has brought growing attention to ethical risks, social impact, and governance issues. COVID-19 research highlights multilingual analysis of public sentiment, health information, and pandemic trends, with increasing focus on vaccine safety, drug efficacy, and mental health.

Conclusions

Overall, text mining research in LIS exhibits several notable trends:

(1) **From intelligence-centered to interdisciplinary integration.** Research has expanded from scientific texts to policy, culture, and health domains, aligning LIS with public administration, digital humanities, and health informatics.

(2) **Large Language Models and Generative AI as new drivers.** These technologies enhance core tasks(Li, Peng & Li, 2024) and introduce new research directions like hallucination detection and content authenticity assessment.

(3) **Text mining and bibliometrics: an evolving synergy.** Their integration enables efficient processing of unstructured data and robust scientific evaluation and trend forecasting (Luo, Lu & He, 2022).

Acknowledgments

This work is supported by Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No.SJCX24_0178).

References

Li, Y., Peng, X., Li, J., Zuo, X., Peng, S., Pei, D., ... & Hong, N. (2024). Relation extraction using large language models: a case study on acupuncture point locations. Journal of the American Medical Informatics Association, 31(11), 2622-2631.

Luo, Z., Lu, W., He, J., & Wang, Y. (2022). Combination of research questions and methods: A new measurement of scientific novelty. Journal of Informetrics, 16(2), 101282.

https://doi.org/10.51408/issi2025_193

Research Collaboration and Leading Role: A Comparative Study on the Academic Communities in Japan and Taiwan

Szu-chia Lo[1], Yuan Sun[2]

[1] *szuchialo@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University (Taiwan)

[2] *yuan@nii.ac.jp*
Information and Society Research Division, National Institute of Informatics (Japan)

Introduction

Collaboration has long been viewed as a preferred strategy for enhancing knowledge or expanding academic research resources, and collaborative approaches to research are encouraged (Katz & Martin, 1997; Kyvik & Reymert, 2017; Ponomariv & Boardman, 2016). Besides to carry out collaborative activities, these topics involved with research collaboration have attracted researchers to study collaboration scenarios (Ponomariov & Boardman, 2016). However, various factors, political policy, health issues and economic stress, which the academic communities might encounter at some points, such as COVID might transform scholarly activities (Melin & Persson, 1996; Ponomariov & Boardman, 2016), and further influence research collaboration. The authors of this work particularly interested in the research collaboration among the universities affiliates and took this as the main theme of this study. Following up the authors' previous studies, the authors continue to take the co-authorship as a flag to present the research collaboration, holding the position of the first or corresponding author was the leading role in the collaboration, and the journal articles were taken as the outputs of the research collaboration. The works were retrieved and examined to investigate the scenario of research collaboration and the role in the collaboration of the universities that are with different research productive strengths in Japan and Taiwan. Considering the long-term features and trends changes, the data of 2014, 2017, 2020 and 2023 were searched for this study and targeted universities were also screened with the idea of "Bradford Law". The study tried to target the following research questions,

-Research productivity of the targeted universities of Japan and Taiwan during the observed period

-Research collaborations of the targeted universities of Japan and Taiwan during the observed period

-Role of the authors in the research collaboration of the targeted universities of Japan and Taiwan during the observed period

Method and data

The authors targeted the journal articles included in Web of Science (WoS) (A&HCI, SCI, SSCI) for this study and the data was retrieved by the names of the affiliations, which were the universities of Taiwan and Japan. The list of universities was obtained from NTU (National Taiwan University) Rankings, and the names of those universities were used for the search. To enclose the trends of research collaboration, the authors kept the articles issued in 2014, 2017, 2020 and 2023 for the study. After the first-round search, the author information was extracted from the bibliographic data, and tagged indicate the authorship, type of collaboration, which was tag either domestic or international collaboration, and the roles of the author in the collaboration, leading role means the author was list either as the first or corresponding author, and supportive was tagged if neither status applied. To investigate the similarity or difference of collaboration of the universities that showed different levels or research strengths, the authors applied the ideas of Bradford Law and grouped the universities into three sets based on the research

productivity, and took the universities listed at the first place of each group for further observation. All the calculations were conducted on a university's basis, and the results were viewed from an institutional perspective. The works done by over 100 universities in Japan and 70 universities in Taiwan, and over 540-thousands articles were examined in this study. Table 1 shows the numbers of universities and articles included in this study

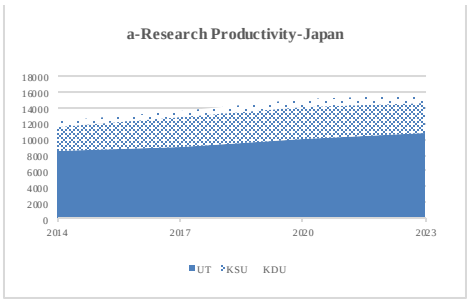
Table 1. Statistics of research data.

	Japan		Taiwan	
	Universities	Articles	Universities	Articles
2014	111	83345	77	37610
2017	112	95477	76	36296
2020	93	105948	42	41223
2023	80	97724	35	42724

Results

With the limited space for the poster, the authors present the research results mainly with figures and related statements.

Major gaps in research productivities were shown among the universities in different productive tiers for both Japan and Taiwan. (Figure 1-a, 1-b)



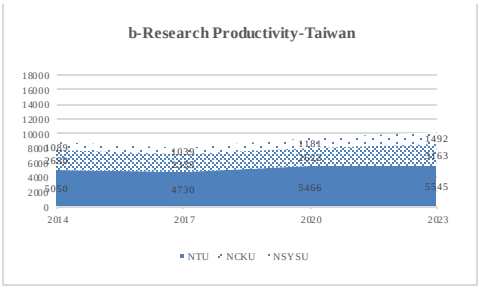
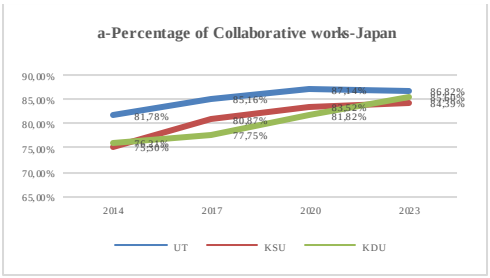


Figure 1. Research productivities of sampled universities: (a) Japan, (b) Taiwan.

The research output relied on highly collaborative efforts, and the dependence continuously increased for both Japan and Taiwan. (Figure 2-a, 2-b) In Japan, the more productive universities gained more chances to collaborate with peers.



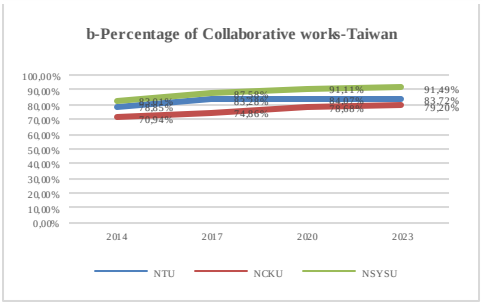


Figure 2. Dependency of research collaboration: (a) Japan, (b) Taiwan.

Domestic collaboration had more advantages during the 2010s, but international collaboration attracted similar efforts after the 2020s.

Sampled universities hold leading roles in close to 50% of the research collaboration with one exception. (Figure 3-a, 3-b). The universities with higher productivity tend to hold the leading position in the collaboration

in Japan, however no major differences are observed in the cases of Taiwan.

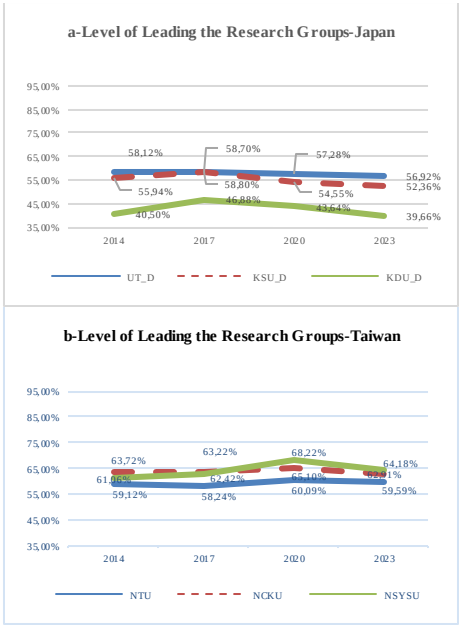


Figure 3. Leading roles in research collaboration: (a) Japan, (b) Taiwan.

Conclusion

The results reflect the trends and the similarities of the strategies taken for the research development of the observed regions. Both domestic and international collaboration gained attention for the sampled universities. The pictures of leading roles in the research collaboration of the academic communities of Japan and Taiwan are a little bit different, the ones with more productive strength had better chances to hold the leading position in the collaboration in Japan, but no strong evidence to show the differences for Taiwan.

Acknowledgments

This work was financially supported by the Universities and Colleges Humanities and Social Sciences Benchmarking Project and the Center for Research in Econometric Theory and Applications which is under the Featured Areas Research Center Program by Higher Education Sprout Project of the Ministry of Education (MOE) in Taiwan.

References

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*, 1-18.

Kyvik, S., & Reymert, I. (2017). Research collaboration in groups and networks: Differences across academic fields. *Scientometrics, 113*, 951-967.

Melin, R, & Persso, G. (1996). Studying research collaboration using co-authorships. *Scientometrics, 36*(3), 363-377.

Ponomariov, B, & Boardman, C. (2016). What is co-authorship? *Scientometrics, 109*, 1939-1963.

https://doi.org/10.51408/issi2025_183

Research on Scientific Frontier Topics Based on Citation Analysis and Content Analysis - Taking the Structural Analysis of Nature Index as an Example

Tan Xiao[1], Li Hui[2], Xu Haiyun[3], Li Jiayu[4], Jin Xiaohong[5], Xi Guiquan[6], Zhang Ting[7], Chen Shu[8]

[1]*tantan46227@163.com*, [2]*tantan83_1@163.com*, [4]*meganli328@hotmail.com*, [5]*531277608@qq.com*, [6]*81289113@qq.com*, [7]*tanx@bjast.ac.cn*, [8]*c.s.luck@163.com*
Institute of Scientific and Technical Information, Beijing Academy of Science and Technology, No.140, Xizhimenwai street, Xicheng District, Beijing,100044 (China)

[3] *xuhaiyunnemo@gmail.com*
School of Management, Shandong University of Technology, No. 266, West Xincun Road, Zhangdian District, Zibo City, Shandong Province, 255000 (China)

Introduction

In the era of global scientific and technological revolution, interdisciplinary integration prevails. Scientific frontier research, emerging after 2005, focuses on identifying technological trends. The Nature Index (NI) measures research output but lacks content analysis. This paper uses NI data, integrating entity-relationship and semantic-structure, to explore frontier topics via text modeling and automatic identification, aiding researchers in spotting hotspots and directions.

Literature Review

Research frontiers, rooted in novel discoveries, exhibit bibliometric patterns like term surges and citation network shifts (Chen, 2009). Detection methods, including Delphi and ML-based ones, fall into citation-clustering and indicator-driven categories.Metrological methods for scientific frontiers rely on high co-occurrence/clustering, using disciplinary knowledge networks (Chen, 2010). Traditional methods lack semantic analysis; this paper introduces semantic-structure approaches (Blei, 2003).

Research Steps for Topic Identification and Structural Detection

This study selects 68 NI journal literatures from 2018-2020, preprocesses keywords, vectorizes text to obtain matrices, constructs an LDA model, builds and standardizes a co-citation matrix, calculates cosine similarity, fuses matrices linearly per Janssens' idea, and uses community-discovery algorithms to identify research frontiers.

Description of the Experimental Dataset Construction

We selected 170,000 papers from 68 NI journals (2018-2020), used citation coupling to build relationships, standardized them, and stored data in SQL Server with ESI_IDs and normalized cocitation degrees.

Construction of Similarity Matrices Based on LDA Modeling

Using LDA model, we map documents to topic vectors, extract keywords from titles/abstracts, apply Gibbs sampling with α=50/K, β=0.01, set K=50 via perplexity, iterate 2000 times, and compute doc similarity by cosine.The topic vector mapped by the article is $d_i = (t_1, t_2, \ldots\ldots, t_K)$, the similarity between two articles is calculated using

$$Sim(d_i, d_j) = \cos\theta = \frac{d_i * d_j}{|d_i| * |d_j|} = \frac{t_1 * t_1' + t_2 * t_2' + \ldots + t_k * t_k'}{\sqrt{t_1^2 + t_2^2 + \ldots + t_k^2} * \sqrt{t_1'^2 + t_2'^2 + \ldots + t_k'^2}}$$

(Formula 1) to obtain the document-document similarity matrix.

Integration of Citation Relationships and Semantic Content

It is necessary to further refine the text knowledge units. The integration method applied in this part is a comprehensive method of topic mining and Fisher relationship fusion algorithm. The integration method applied in this part is a comprehensive method of topic mining and Fisher relationship fusion algorithm.

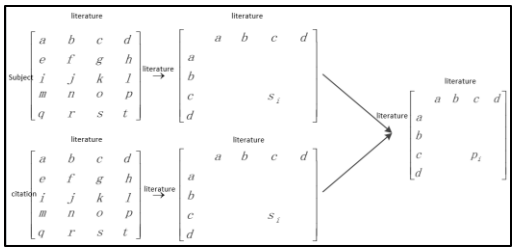


Figure 1. Framework of the Relationship Fusion Method Based on Fisher.

according to the semantic information, the correlation relationships in the network will be adjusted, and their weights will be adjusted:

$$Sim(i,j) = \begin{cases} \lambda \cdot Sim_{LDA} + (1-\lambda) \cdot cite_{couple}, \text{cite}(i,j) = 0 \text{ and } sim_{LDA}(i,j) > a; \text{cite}(i,j) > 0 \\ \quad \text{and } sim_{LDA} \geq b \quad \lambda \epsilon (0,1) \\ 0, cite(i,j) > 0 \text{ and } sim_{LDA}(i,j) < b\ ;\ \text{cite}(i,j) = 0 \text{ and } sim_{LDA}(i,j) \leq a \end{cases}$$

(Formula 2)

Community Detection and Topic Extraction

Community detection differs from traditional clustering by focusing on network node-link relationships. This study fused standardized citation networks and LDA-based similarity, then applied the Louvain algorithm—optimizing modularity via iterative community division and aggregation—for efficient large-scale community detection.

For the literature related to NI from 2018 to 2020, the steps in 3.3 were used for analysis, and secondary clustering of the communities was performed to form the following figure.

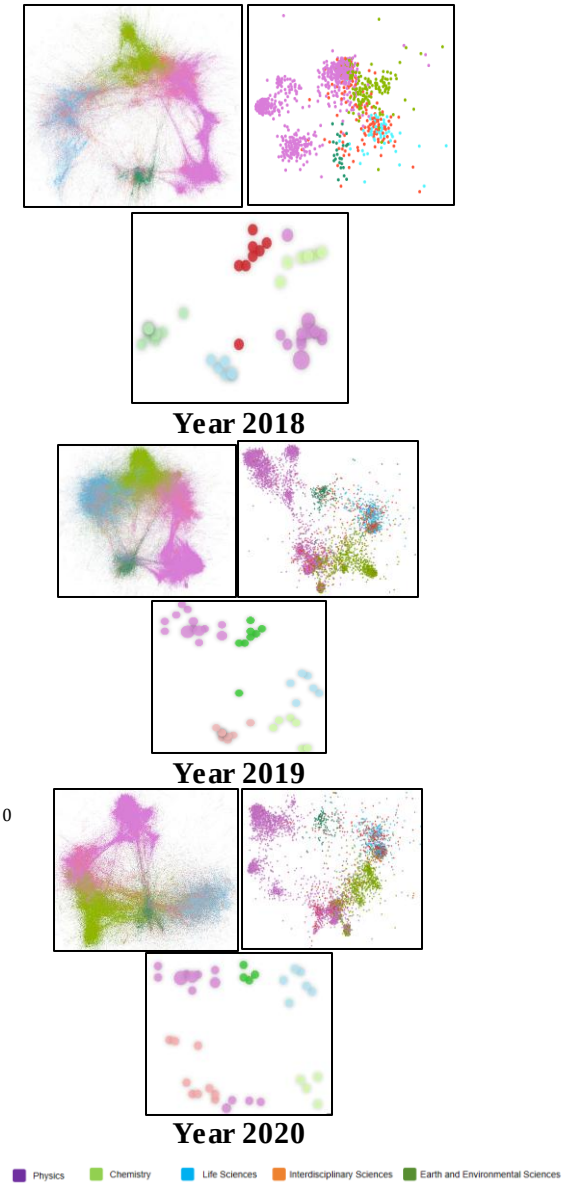


Figure 2. Clustering Diagram Formed by Applying Community Detection Algorithm from 2018 to 2020.

Each circle represents a research topic and consists of several pieces of literature, and its size is positively correlated with the number of documents that constitute the community. In the clusters formed by re-clustering, the keywords representing each cluster are extracted to form the current domain topics and the cross-topic relationships between disciplinary fields.

From an overall analysis, it is found from the community clustering diagrams of the three years from 2018 to 2020 that the highly

interdisciplinary research fields are showing an upward trend; the interdisciplinary integration is becoming wider and wider, showing a spreading trend.

Conclusion

This paper detects frontier field dimensions via "entity-relationship" and "semantics-structure" integration, using 3-year NI journal data with citation-topic combination to identify frontier topics, aiding interdisciplinary research. Limited by space, it only shows overall structural characteristics. Future research will incorporate diverse S&T documents (plans, patents, projects) to construct dynamic knowledge networks, integrate multi-entity relationships, strengthen entity attributes, classify frontier topic types, and enhance identification accuracy by linking topics to classification attributes.

Acknowledgments

This article is the outcome of the projects,"Research on the emergence mechanism and identification of disruptive technologies from the perspective of innovation ecological network" (No.9242006) supported by Beijing Natural Science Foundation, "Early Recognition Method of Transformative Scientific and Technological Innovation Topics based on Weak Signal Temporal Network Evolution analysis" (No.72274113) supported by the National Natural Science Foundation of China and the Taishan Scholar Foundation of Shandong province of China (tsqn202103069).

References

Chen, S. J. (2009). Survey of Approaches to Research Front Detection. *New Technology of Library and Information Service*, 28 - 33.

Chen, L. J. (2010). Research on the Three Evolution Stages of Knowledge Linking Theory and Practice. *Library and Information Service*, 54(12), 46 - 49+63.

Wang, W. B., Cheng, H. M., & Wang, L. J. (2013). Analysis of the Research Situation of China's Strategic Emerging Industries Based on Co-word Analysis. *Science & Technology Progress and Policy*, 30(21), 57 - 60.

Li, G., & Li, Y. (2011). A Co-word Analysis Method Based on Keyword Weighting. *Information Science*, 29(03), 321 - 324+332.

Bai, R. J., Zhang, Y. H., Zhang, Y. J., Ju, Z. H., & Feng, M. Y. (2024). Research on the Identification of Interdisciplinary Frontiers Based on Dual Measurement of Citation and Theme. *Journal of Modern Information*, 44(10), 27 - 40+63.

https://doi.org/10.51408/issi2025_206

Research on the Path of China's Construction of a World Science and Technology Power

Wang Kaile[1], Chen Yunwei[2]

[1]*wangkl@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, No.289, QunXian Nanjie, Tianfu New Area, Chengdu (China)

[2]*chenyw@clas.ac.cn*
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, No. 1 Yanqi Lake East Road, Huairou District, Beijing (China)

Introduction

In recent years, China's technological development has entered a phase of accelerated progress, achieving remarkable results that have garnered global attention. "Breakthroughs in fundamental and frontier research, significant advancements in strategic high-tech fields, the effectiveness of innovation-driven high-quality development, breakthroughs in reforming the science and technology system, and progress in international cooperation have laid a solid foundation for building a strong scientific and technological nation." A December 2023 editorial in Nature highlighted that global science is increasingly dividing into two parallel systems—one centered on North America and Europe, and the other centered on China. Similarly, a September 2024 news analysis in Science pointed out that Chinese scientists frequently cite domestic research, which may distort China's ranking in global research metrics. In June 2024, The Economist featured a cover story titled "The Rise of Chinese Science: Welcome or Worrying?" and published an article, China has become a scientific superpower, sparking widespread global debate. The article used two key sets of data—China's output of high-quality scientific papers and its dominance in the Nature Index, metrics highly regarded by Western nations—as evidence that China has surpassed the U.S. and Europe in these areas. It further emphasized China's massive investments in talent, funding, and infrastructure, positioning it as a global scientific superpower. However, does this truly indicate that China's scientific and technological strength has surpassed that of Western scientific powerhouses and now occupies a globally leading position? This question demands objective analysis and cautious evaluation to ensure an accurate understanding of China's status in global science and technology.

The distance between China's path to becoming a technological powerhouse

This part compares the gaps between China and technologically advanced countries from multiple dimensions. While China's investment in basic research has shown consistent growth, the overall foundation remains relatively weak. In terms of absolute expenditure, China's basic research spending reached $460 billion in 2021. However, this is significantly lower than the United States, whose spending on basic research amounted to $1.2 trillion—nearly three times that of China. As of 2023, Chinese applicants have ranked first globally for five consecutive years in international patent applications filed through the Patent Cooperation Treaty (PCT). However, when evaluated through indicators that better reflect technological content and economic value, the shortcomings of Chinese patents become evident. As shown in Table 1, China lags far behind the United States in terms of the number of high-quality patents with more than 100 claims or a patent value exceeding $1 million. China's number of winners in multiple international authoritative awards is only in single digits, with a huge gap compared to the United

States, and its academic influence in the international arena is relatively small.

Table 1. Total number of patents with different claims and patent values in China and the United States.

Country of Origin	Claims > 10 (sets)	Claims > 50 (sets)	Claims > 100 (sets)
China	8306736	20423	2,850
United States	1939091	84485	14,789
Country of Origin	**Patent Value > $10,000 (sets)**	**Patent Value > $100,000 (sets)**	**Patent Value > $1,000,000 (sets)**
China	2994695	1331175	61273
United States	1435166	551614	120978

In 2022, the R&D investment intensity of Chinese enterprises accounted for 1.98% of GDP, significantly lower than that of the United States (2.83%), Germany (2.11%), and Japan (2.70%). This highlights the relatively weak overall innovation capacity of Chinese enterprises. Between 2014 and 2023, the only countries surpassing China in the number of highly cited researchers globally were the United States, the United Kingdom, and Germany. Since 2019, however, China has firmly held second place. Despite this steady growth, the number of highly cited scientists in China remains significantly lower than in the United States, amounting to less than half of the U.S. total (Figure 1).

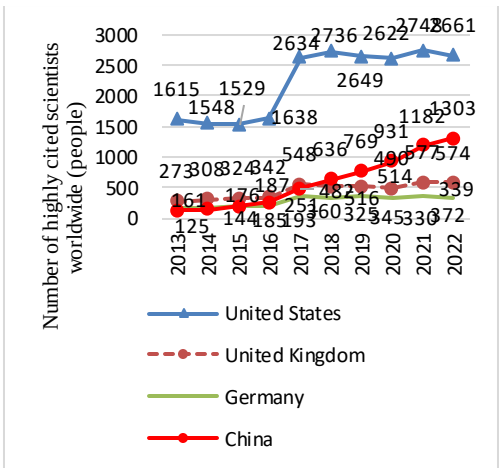


Figure 1. Evolution of the number of highly cited scientists worldwide from 2014 to 2023.

Exploring the Path to Becoming a Global Science and Technology Power

Amid profound changes unseen in a century, a new wave of scientific and technological revolution and industrial transformation is reshaping the global landscape, with China steadily moving toward the center of the world stage. In response to the urgent need to build itself into a global science and technology power, China's current development pathway requires significant optimization and improvement. Against this backdrop, determining the direction and strategy for China's science and technology advancement has become a pressing priority. First is advancing strategic-oriented basic research and frontier-oriented exploratory basic research. Second is supporting enterprise-led innovation consortia to enhance the efficiency of technology commercialization. Third is improving policies for attracting and cultivating high-Level innovation talent.

References

Xi J.P. (2024). Speech at the National Science and Technology Conference, National Science and Technology Awards Conference, and Academician Conference of the Chinese Academy of Sciences and Chinese Academy of Engineering. https://www.gov.cn/yaowen/liebiao/202406/content_6959120.htm.

Song D.C, Xiao S, Li T.M, et al. (2024). Comparison of open sharing modes of foreign large scale scientific facilities and implications for China. *Bulletin of Chinese Academy of Sciences*, 39(03):447-458.

Jiefang Daily. (2023). Research report on artificial intelligence big model maps released, China ranks second in the world in terms of the number of big models. https://www.shanghai.gov.cn/nw4411/20230529/ea872bcdea80457ca2d474a33a39c9ad.html.

SOHU. (2024). China has been ranked first in global new energy vehicle production and sales for 9 consecutive years. https://mil.sohu.com/a/771551196_330740.

China News. (2024). Academician XueQikun: China's quantum technology is in the world's top tier. https://www.chinanews.com.cn/gn/2024/06-24/10239309.shtml.

People's Daily Online. (2023). The full text is here! The President Xi Jinping delivers a New Year's message for 2023. https://baijiahao.baidu.com/s?id=1753855432493893844&wfr=spider&for=pc.

Guangming Net. (2023). China's Science and Technology Innovation Brings Benefits to the World - China's International Science and Technology Cooperation Achieves Abundant Achievements in 2022. https://m.gmw.cn/baijia/2023-01/16/36306263.html.

Xinhuanet. (2023). Building an open innovation ecosystem with global competitiveness. http://www.xinhuanet.com/politics/20230608/7ffc49d5cf0744c59c8e5954c15fe7da/c.html.

The People's Bank of China. (2023). Central Bank Report: The consumption rate of Chinese residents is still significantly lower than that of high-income countries. http://www.ce.cn/xwzx/gnsz/gdxw/202302/25/t20230225_38412801.shtml.

Li H.G. (2016). Innovation culture is an important element of technological innovation. *The People's Daily*，2016(5).

https://doi.org/10.51408/issi2025_197

Revolutionizing Medical Processes Through Phygital Technology: a Multiple Case Study Approach

Eugenio Oropallo[1M], Cinzia Daraio[2], Simone Di Leo[3], Fabio Nonino[4]

[1]*eugenio.oropallo@uniroma1.it*, [2]*daraio@diag.uniroma1.it*, [3]*dileo@diag.uniroma1.it*, [4]*fabio.nonino@uniroma1.it*

Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), Sapienza University of Rome, Via Ariosto 25, Rome, 00185 (Italy)

Introduction

In our increasingly interconnected world, the line between the physical and digital is blurring: Phygital technology, the synergistic fusion of these two realms, is transforming industries, particularly healthcare, by opening new frontiers in how we interact with the world. Phygital technology seamlessly integrates physical and digital elements to create engaging, interactive, and personalized experiences (Kalra et al., 2023). Healthcare, inherently focused on human interaction and personalized care, faces unique digital-age challenges (Frascolla, 2020). Emerging technologies in the healthcare field, within the digital and virtual ecosystems, can facilitate hyper-personalized, data-driven care, leading to earlier disease detection, tailored therapies, and improved patient outcomes through individualized, predictive, and empathetic engagement (Paton et al., 2024). Phygital technology offers a powerful solution in this direction, promising to enhance the patient experience, streamline healthcare processes, and revolutionize diagnosis, treatment, and care. It leverages the power of the digital to enrich and amplify real-world interactions. This integration rests on three key pillars (Maci, 2024):

- Physical-Digital Integration: Phygital technology bridges the gap between the physical and digital using sensors, smart devices, and digital platforms. These interconnected elements create a dynamic ecosystem where information flows effortlessly between the real and virtual.
- Interactivity and Personalization: Phygital technology enables highly interactive and personalized experiences, tailoring the user journey to individual needs and preferences. This interactivity can be achieved through touchscreens, augmented reality, virtual reality, and other human-machine interfaces.
- Data Collection and Processing: Phygital technology facilitates the collection and analysis of vast amounts of data from diverse sources. This data provides valuable insights, informs better decision-making, and further personalizes the user experience.

Phygital technology has enormous potential to revolutionize healthcare, with applications spanning from patient management to cutting-edge medical research. It is possible to consider some examples; Phygital solutions empower patients through personalized health management apps, remote vital sign monitoring, and convenient telemedicine consultations. Besides, Phygital technology optimizes workflows through digital medical record management, efficient appointment scheduling, and streamlined administrative tasks; specifically Phygital tools support more accurate diagnoses and effective treatments through medical image analysis, surgical simulations, and robotic-assisted procedures. Moreover, Phygital technology accelerates research by enabling the collection and analysis of massive datasets, the creation of virtual organ and tissue models, and the simulation of complex clinical scenarios. In this way, Phygital technology represents a promising frontier for healthcare: its ability to merge the physical and digital worlds offers a unique opportunity to improve patient care, optimize healthcare delivery, and drive innovation in medical research. Drawing on the research conducted into the "Phygital Twin Technologies for innovative Surgical Training & Planning" of the Rome Technopole program, this poster presents a first artifact solution related to the phygital

technology was analysed, allowing to investigate the cross-advantages of a physical product with its digital twin. For this type of artifact, it will be necessary to understand how to guarantee the immediacy, immersion and interaction of users with respect to the clinical case for which the phygital experience is designed.

Data & Method

A multiple case study was conducted with experts of this technology at the Sapienza, University of Rome. This analysis informed the development of a framework demonstrating what features the phygital phantom product needs for its architecture in order to be a valid product that can be used not only in the real world and offline mood but also in an online integration perspective in a metaverse world or on-line mood. A Design Science approach guided our analysis. Design Science approach is a problem-solving approach that involves designing, implementing, and evaluating artifacts to address specific needs, bridging the gap between research and practice (Tarpey and Mullarkey, 2021; Guggenberger et al., 2020). Our research followed the Design Science approach process, including defining the ecosystem and problem space, which involves identifying key actors (e.g., individuals and organizations), applications, and goals (Centobelli et al. 2021; Cerchione et al., 2022). A well-defined ecosystem enhances the relevance of the resulting artifact. Our knowledge base, comprising established methodologies and theoretical foundations, provided the necessary research context, supporting the qualitative findings and enhancing the quality and effectiveness of the framework found (Hevner et al., 2004). The framework's value is assessed based on its relevance to real-world business requirements.

To collect useful information to define the ecosystem and basic knowledge, data collections were carried out using a mixed methodology, the Analytical Hierarchy Process (AHP) and Fuzzy set theory (FST). Specifically, we were asked to evaluate different aspects that emerged in the literature, with the possibility of adding others related to the interviewees' experience. In this way, it was possible to dynamically define and evaluate all the aspects that emerged from the surveys carried out. Using a mixed methodology, AHP/FST allows us to integrate additional variables during the process without losing the data collection that was previously carried out. The old respondents had to answer only the questions relating to the new variables without returning to the answers already given.

Results and conclusions

The resulting framework offers an innovative approach to integrating diverse medical requirements, coordinating stakeholders and activities, and seamlessly connecting with existing systems to automate and optimize both local and remote healthcare workflows. Key benefits of implementing the Phygital technology in healthcare include personalized health data tracking, advanced analysis of patient clinical data, and the potential for eliminating paper-based medical records. Furthermore, training activities, users' medical skills and knowledge diffusion of healthcare best practices result the main advantages that this technology can offer to the ecosystem where it is adopted. However, challenges remain (Paquin et al., 2023), particularly regarding the management of sensitive data flows and the potential risks to user privacy and ethical considerations (Koohsari et al., 2023). Successfully integrating the Phygital technology into healthcare will require addressing these challenges and ensuring secure and ethical interactions between doctors, patients, and devices.

Acknowledgments

The poster was partially supported by project TECHNOPOLE - Flagship 4; Rome_Tech Spoke_2_DIAG, CUP B83C22002820006 and the "SEcurity and RIghts In the CyberSpace (SERICS) (ECS – Rome Technopole)".

References

Centobelli, P., Cerchione, R., Del Vecchio, P., Oropallo, E. and Secundo, G. 2021. Blockchain technology design in accounting: Game changer to tackle fraud or technological fairy tale?, Accounting,

Auditing, Accountability J., 35(7), pp.1566–1597.

Cerchione, R., Centobelli, P., Riccio, E., Abbate, S., and Oropallo, E. 2022. Blockchain's coming to hospital to digitalise healthcare services: Designing a distributed electronic health record ecosystem, Technovation, 15, 102480.

Frascolla, V. (2020). Un mondo sempre più phygital: L'impatto dei Digital Signage sulla willingness to pay a higher price del consumatore. LUISS - Dipartimento di Impresa e Management.

Guggenberger, T., Schweizer, A., Urbach, N. 2020. Improving interorganizational information sharing for vendor managed inventory: Toward a decentralised information hub using blockchain technology, IEEE Trans. Eng. Manage. 67(4), pp. 1074–1085.

Hevner, A.R., March, S.T., Park, J. and Ram, S. 2004. Design science in information systems research, MIS Quart., 28(1), 1.

Kalra, S., Tiwaskar, M., Shrestha, D., Somasundaram, N., & Gokhalay, S. (2023). Digital Nerve Care Forum: Innovative Healthcare Professionals Education on Neuropathy. Journal of Association of Physicians of India, 79(10), 89–92. Scopus. https://doi.org/10.59556/japi.71.0346

Koohsari, M.J., McCormack, G.R., Nakaya, T., Yasunaga, A., Fuller, D., Nagai, Y., Oka, K. 2023. The Metaverse, the built environment, and public health: opportunities and uncertainties, Journal of Medical Internet Research, 25: e43549.

Maci, L. (2024). Phygital: Cos'è e perché funziona bene con l'industria culturale.

https://www.economyup.it/innovazione/phygital-cose-come-funziona-e-come-sfruttarlo-per-migliorare-la-customer-experience/

Paquin, V., Ferrari, M., Sekhon, H., Rej, S. 2023. Time to think "Meta": a critical viewpoint on the risks and benefits of virtual worlds for mental health, JMIR Serious Games, 11:e43388.

Paton, C., Borycki, E. M., Warren, J., Kushniruk, A. W., & English, M. (2024). HCI-modelling for improving the clinical usability of digital health technologies. Methods, 227, 60–77. Scopus. https://doi.org/10.1016/j.ymeth.2024.04.019

Tarpey, R. and Mullarkey, M. 2021. Engineering innovative clinical resource management by design: A guided emergent search through a complex adaptive systemof systems," IEEE Trans. Eng. Manage. doi: 10.1109/TEM.2021.3059590.

https://doi.org/10.51408/issi2025_181

Single journal bibliometric case studies

Ilya Gorelskiy[1], Daniel Karabekyan[2], Alexander Karpov[3]

[1]*igorelskiy@hse.ru*, [2]*dkarabekyan@hse.ru*, [3]*akarpov@hse.ru*
HSE University, Moscow (Russia)

Abstract

This study presents a comprehensive analysis of single journal bibliometric self-studies — bibliometric case studies published in the same journal they analyze — to explore their evolution, impact, and emerging ethical challenges. This work uses the OpenAlex database to identify 643 self-studies from 565 journals (1988–2024), offering the largest quantitative examination to date. Our methodology combines keyword filtering (e.g., journal titles in article titles, terms like "bibliom*" in abstracts) with manual validation to exclude non-relevant content (e.g., editorials, thematic subsets), ensuring a focused dataset. A key finding is the rising trend of self-studies authored by professional bibliometricians unaffiliated with the journal's core community, particularly post-2020. These externally produced papers, frequently published in high-impact journals, yield mutual benefits: authors gain visibility in prestigious venues, while journals enhance their citation metrics. Our findings show a dual reality: single journal self-studies offer valuable field-specific insights but are increasingly exploited for bibliometric gaming.

Introduction

Scientific journals serve as homogeneous collections of research output, united by shared disciplines, editorial policies, and publishing standards. These collections are critical for bibliometric analysis, particularly single journal bibliometric case studies, which provide insights into the intellectual evolution, editorial practices, and citation dynamics of individual journals. Such studies are often published within the analyzed journal itself, termed here as single journal bibliometric case studies or self-studies. While prior surveys by Tiew (1997) and Anyi, Zainab & Anuar (2009) categorized these studies qualitatively using small samples (102 and 82 papers, respectively), their analyses focused on periods ending in 2008, leaving recent trends underexplored.

Single journal self-studies vary in scope: some trace a journal's history (e.g., Arrow et al., 2011; Margo, 2011), others evaluate editorial performance (e.g., Zink, 1950), and many focus on citation-based bibliometrics. Historically, such studies were authored by members of the journal's community. However, recent years have seen a rise in contributions from external bibliometricians, raising questions about motivations and ethical practices. This study addresses these gaps by analyzing the largest dataset of single journal self-studies to date (1988–2024), examining their evolution, impact, and emerging ethical challenges.

Methodology

We extracted data from OpenAlex using a search strategy targeting papers with journal titles in their article titles (including full titles, abbreviations, and variants which are available in OpenAlex). From an initial pool of 27 484 papers, we applied inclusion criteria: (1) keyword filtering ("bibliom*" or "scientom*" in abstracts) (1147 left after filtering); and (3) manual validation to exclude papers analyzing thematic subsets of a journal's output or several journals at the same time (e.g., Skop, Tonyan & Cassiday, 2019). The final dataset comprises 643 self-studies from 565 journals. OpenAlex work_ids can be provided.

Results and Dataset Overview

The 643 self-studies span 1988–2024, with a sharp increase in 2018 (more than 66.7% of the papers were published from 2018 to 2024). Distribution is given in Figure 1.

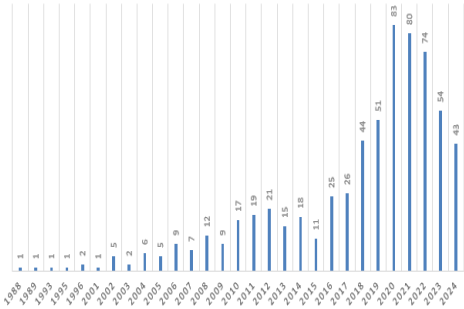


Figure 1. Number of papers in the dataset for a given year.

There are 58 journals that have published more than 1 self-study. In Table 1 journals with 3 and more self-studies are given.

Self-studies are more popular among Social Science journals. At Figure 2 shares of different fields are given. Note, that two or more fields can be attributed to one publication.

Table 1. Journals with the highest number of self-studies.

Journal name	*papers*
Australasian Journal of Educational Technology	6
Scientometrics	6
Information Sciences	5
Journal of Business Research	3
Journal of Craniofacial Surgery	3
Journal of Cross-Cultural Psychology	3
Journal of Product Innovation Management	3
Library philosophy and practice	3
Medicine	3
Naunyn-Schmiedebergs Archives of Pharmacology	3
Retos	3

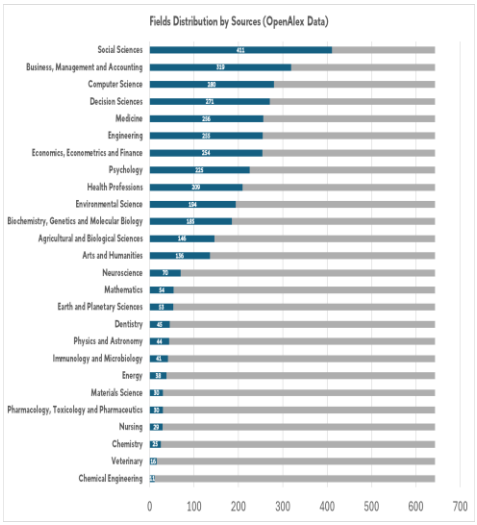


Figure 2. Fields Distribution.

Single journal bibliometric self-studies have significant influence. The average number of citations is 24 while the median is 5. The most cited paper in our database (Ramos-Rodríguez & Ruíz-Navarro 2004) has 1395 citations. Usually, single journal bibliometric self-studies provide general overview of main trends in research field, comparison with related research areas. Because of deliberate development of science, some findings stay relevant for a long time.

Historically, such studies were authored by members of the journal's community. With advances in bibliometric research, development of bibliometric instruments, and scholars' engagement with bibliometric indicators the number of single journal case studies is increasing. The more interesting trend is that such papers are written by professional bibliometricians that do not belong to the journal's scientific community. The three most productive coauthors published 51, 29, and 22 correspondingly. Most of these papers are published between 2020 and 2024. Many of these papers have very good citation performance. They cite their own related research in different journals. Both journals and authors win from such strategy.

We have found only 79 papers that are written by coauthors of at least 10 single journal bibliometric self-studies. The share of these papers has significantly increased over the last 5 years, most papers are still authored by

scholars with relatively small number of single journal bibliometric self-studies.

Discussion and conclusion

Single journal bibliometric self-studies serve dual roles: they provide valuable syntheses of disciplinary progress but are increasingly exploited for bibliometric gaming. While most studies remain ethically sound, the rise of templated papers highlights vulnerabilities in current bibliometric and editorial systems. Journals benefit from heightened visibility through these studies, yet risk enabling manipulative practices that distort impact metrics.

This research is still in progress. Future research should explore longitudinal citation patterns of self-studies and develop frameworks to balance their academic value with ethical safeguards.

References

Anyi, K.W.U., Zainab, A.N. & Anuar N.B. (2009). Bibliometric studies on single journals: a review. Malaysian Journal of Library & Information Science, 14(1), 17-55.

Arrow, K.J., Bernheim, B.D., Feldstein, M.S., McFadden, D.L., Poterba, J.M. & Solow, R.M. (2011). 100 years of the "American Economic Review": The top 20 articles. American Economic Review, 101(1), 1–8.

Margo, R.A. (2011). The economic history of the "American Economic Review": A century's explosion of economics research. American Economic Review, 101(1), 9–35.

Ramos-Rodríguez, A.R. & Ruíz-Navarro, J. Changes in the intellectual structure of strategic management research: a bibliometric study of the Strategic Management Journal, 1980–2000. Strategic Management Journal, 25(10), 981-1004.

Skop, E., Tonyan, J. & Cassiday A. (2019). Considering Refugees Through 100 Years of Geographical Review. Geographical Review, 109(4), 598-614.

Tiew, W.S. (1997). Single journal bibliometric studies: a review. Malaysian Journal of Library & Information Science, 2(2), 93-114.

Zink, H. (1950). The growth of the American Political Science Review, 1926-1949. American Political Science Review, 44(2), 257–265.

https://doi.org/10.51408/issi2025_175

Stop, little pot! Are there too many scientometric studies?

Ekaterina Dyachenko[1], Alexey Zheleznov[2], Maxim Dmitriev[3], Katerina Guba[4]

[1] *edyachenko@eu.spb.ru*, [2]*azheleznov@eu.spb.ru*, [3]*mdmitriev@eu.spb.ru*, [4]*kguba@eu.spb.ru*
Center for Institutional Analysis of Science and Education, European University at Saint Petersburg, Gagarinskaya 6/1A Saint Petersburg 191178 (Russian Federation)

Introduction

Scientific research is growing rapidly, with an exponential rise in published studies (Bornmann & Mutz, 2015). Many scientometric studies aim to help scientists cope with a vast amount of publications. Scientometric methods assist in identifying the structure of the scientific field and trends in its development, thereby help scientists in structuring their attention. Are they good at it? Or does scientometrics mainly serve its own purposes? This study examines scientometric research on climate science. Gerald Stanhill (2001) was among the first to show the exponential growth of climate studies, echoing de Solla Price's (1963) broader observations. As climate research has surged, scientometric studies on the topic have also grown, especially in recent years. This raises a question: do these studies genuinely support climate scientists?

We address the following research questions:
(A) whether scientometric studies of climate research field are useful for the climate researchers;
(B) whether the growth of scientometric studies of climate research is accompanied by an increase in the diversity of research questions, methods, and objects studied;
(C) is there a difference in the level of attention climate scientists give to scientometric studies that use simple questions and methods compared to those employing more complex approaches?

Methodology

To answer question (A) we applied bibliometric analysis looking at how scientometric studies on climate research are cited in the climate studies. For question (B) we used the qualitative text analysis based on extracting the elements of content of the papers. To answer question (C) we used the data obtained on the previous steps and introduced the category 'the paper with basic analysis' for scientometric studies.

To obtain the set of papers with scientometric studies of climate research we used Scopus database and searched for the combination or terms in title, abstract and keywords of the documents. We limited the search to document types "article" and "review". The exact query was the following:

TITLE-ABS-KEY (climate AND (bibliometr* OR scientometri*)) AND (DOCTYPE (ar) OR DOCTYPE (re)).

The search query returned 1441 results published between 1996 and 2023. We manually checked all the articles to screen out false positives. For these papers we obtained the citation indicators provided by SciVal, as well as the indicators of the journals. We also stored metadata of all papers citing the articles from our original dataset according Scopus database.

To investigate the evolution of scientometric studies of climate research we conducted content analysis of the full texts of the papers. We coded the following characteristics of each scientometric study: software used in analysis, methods of analysis, database(s) used, units of analysis (countries, journals, authors, fields, topics, organizations, etc.), and others. We introduced a "basic analysis" category assigned to papers from the dataset based on the above data. We defined 'the paper with basic analysis' as descriptive study which could be performed almost entirely with functionality of academic databases and/or VOSviewer software. More specifically, we consider the following as the elements of basic scientometric analysis: (a) yearly dynamics of the number of publications, (b) top authors, most publishing countries or organizations, top journals, keywords, subject categories, (c) share of publications with international co-authorship, top partnering countries and institutions, (d) VOSviewer maps (terms, authors, organizations, articles). According to

our definition, the study with basic scientometric analysis contains some combination of those elements, and it does not contain other types of scientometric analysis.

Results

The first scientometric study of climate research – at least the first one we obtained from Scopus – was published in 2001. Until 2014 such studies were sporadic with no more than five articles each year. In the most recent years the number of such studies became surprisingly high – more than hundred articles annually. The geography of scientometric studies of climate research is quite diverse with 97 countries present in the affiliations of the authors. China is the undisputed leader, 38% of papers have at least one co-author from this country. The USA and Spain are the next most active with 8% of papers.

The set of journals where the scientometric studies were published is also diverse. It includes 310 journals, but only fraction of them publish such studies regularly (13 journals published 10+ papers). The two journals with the biggest number of papers – *Sustainability (61 papers*) and *Environmental Science and Pollution Research* (49 papers) – both have a controversial reputation. *Sustainability* is published by MDPI, the publisher with questionable quality standards (Oviedo-García, 2021). *Environmental Science and Pollution Research* published by Springer was recently put into warning list in Clarivate databases due to suspicious citation patterns. The fact that these two journals published scientometric studies on climate most actively in the recent years suggests that some scientometric research is done for the sake of publication itself. The share of articles in high impact journals (journals with high SJR indicatior) decreases as well as the share of papers cited above the global average (paper with FWCI > 1). This shows that on average the impact of scientometric studies of climate research declines.

To look at this explosive growth of scientometric studies from another perspective we analyzed the content of the articles and investigated whether they were becoming more diverse, extensive and sophisticated. We omit the part of the results here because of the space limitation and include only the Figure 1 which shows how prevalent were papers with basic scientometric analysis throughout the period covered. We see such studies do not constitute the majority of all scientometric studies, with the share around 30-40% in the recent years. Apart from the share, the number of such studies has been growing, and recently there were several dozen published each year. We also aimed to explore the origins of studies with basic analysis. We wanted to know whether there is a significant imbalance in which part of the world these studies come from, and whether there is any discernible trend. Preliminary results show that both parts the Global North and the Global South actively produce studies with basic scientometric analysis, but most of the "simple" studies have authors from the Global North.

In our interest to "papers with basic analysis" there is no premise that such studies are not valuable. Some of the papers with basic analysis were done by the most reputable scientometric experts (for example, Haunschild, Bornmann & Marx, 2016), and judging by the number of citations these studies are highly valued. According the data from SciVal the average Field-Weighted Citation Impact for scientometric studies of climate research is 1.52. For papers with basic analysis the average is even higher than for the rest of the studies (1.76 vs. 1.41).

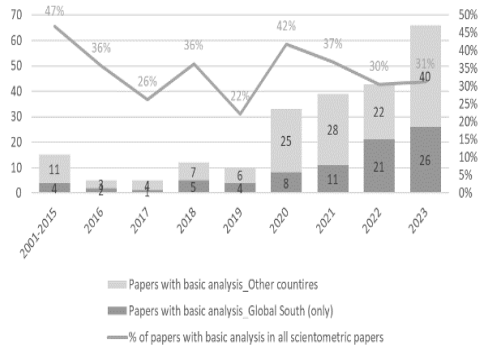


Figure 1. Prevalence of studies with basic scientometric analysis among scientometric studies of climate research.

Respected experts on scientometrics discussed the crisis in the field even before it began to grow explosively – it seems their warnings proved prophetic (Glänzel & Schoepflin, 1994). Today, many scientists are motivated to produce articles not solely by epistemic motives or the desire to attract attention, but also by the pragmatic motives.

Bibliometrics enables a large number of scientists to produce articles with a relatively low threshold of entry into the topic. We found that the rapid growth of such studies is partly due to the production of studies with basic analysis, on the one hand, and, on the other hand, to publications in low-tier journals. At the same time, we see that the citation rates of scientometric studies are on average at a decent level, including studies with basic analysis. The analysis of who cites scientometric studies (not described in detail above) showed that mostly citations are made in non-scientometric papers. Thus, scientometric analysis is clearly useful for scientists in other fields to understand the structure of the literature field.

Acknowledgments

The study of "papers with basic analysis" was funded for E.D. and M.D. by a grant from the Russian Science Foundation № 25-28-01490

References

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11), 2215-2222.

Glänzel, W., & Schoepflin, U. (1994). Little scientometrics, big scientometrics... and beyond?. *Scientometrics*, 30, 375-384.

Haunschild, R., Bornmann, L., & Marx, W. (2016). Climate change research in view of bibliometrics. *PloS one*, 11(7), e0160393.

Oviedo-García, M. Á. (2021). Journal citation reports and the definition of a predatory journal: The case of the Multidisciplinary Digital Publishing Institute (MDPI). *Research evaluation*, 30(3), 405-419a.

https://doi.org/10.51408/issi2025_191

The Effectiveness of Large Language Models in Predicting User Question Preferences on ResearchGate Q&A

Lei Li[1], Yue Hu[2], Hui Peng[3]

[1]leili@bnu.edu.cn,[2]202221260043@mail.bnu.edu.cn,[3]hui_peng@mail.bnu.edu.cn
Department of Information Management, School of government, Beijing Normal University, Beijing (China)

Introduction

In the digital era, academic social platforms such as ResearchGate have become crucial venues for experts, scholars, and researchers across various fields to pose academic questions and receive high-quality answers. Therefore, predicting public preferences for academic questions can help platforms recommend content more accurately, enhance user experience, and assist researchers in understanding current research trends, thereby promoting academic exchange and development. In recent years, methods for predicting user preferences for online-generated content have primarily relied on machine learning algorithms based on feature engineering (Liao et al., 2019), which require high accuracy in feature selection and have limitations in terms of method portability and prediction accuracy. The rapid development of large language models (LLMs) has driven their widespread application across various domains. One of the most promising areas is the use of LLMs for text comprehension and assessment, commonly referred to as "LLMs-as-judges" (Li et al., 2024). This study essentially leverages LLMs to evaluate the popularity of academic questions. The advantages of LLMs in text understanding and assessment provide new possibilities for predicting user preferences for academic questions, offering the potential to reduce excessive reliance on manually crafted features and improve prediction accuracy.

Moreover, unlike general social media platforms, academic Q&A websites place greater emphasis on gaining inspiration and acquiring knowledge of interest through questions. As a result, user question preferences largely depend on the content itself. Leveraging LLMs to predict question popularity by deeply understanding and extracting insights from question content may yield better results. Therefore, this study collects question data from multiple disciplines on ResearchGate Q&A, processes the semantic information in question texts using LLMs, and employs fine-tuning techniques to build a predictive model for user academic question preferences. This approach aims to reveal the potential applications of LLMs in this evaluation task.

Data collection

On the ResearchGate Q&A platform, questioners typically add multiple topic tags to their questions to attract scholars with similar research interests to participate in discussions. This study selected ten specific academic topics from the platform's popular themes, ensuring comprehensive coverage of the five major subject categories in the Web of Science (WOS). Additionally, the broad topics of "learning" and "scientific research" were included to ensure diversity in question types and sufficient data volume, enabling a comprehensive evaluation of LLMs in predicting user academic question preferences.

A Python web crawler was used to collect all questions under these twelve topics, including details such as question titles, descriptions, posting times, view counts, follow counts, answer counts, and recommendation counts. From this dataset, 10,000 questions were randomly sampled for analysis, with the number of questions for each of the 12 topics shown in Table 1.

Table 1. Number of Questions in Each Topic.

Topic	No.	Topic	No.
Molecular Biology	600	Computer Science	111 7
Ethics	404	English	337
Chemistry	170 7	Learning	644
Philosophy	132 2	Journalism	130 8
Economics	336	Social Sciences	411
Psychology	568	Scientific Research	124 6

Model training

Based on mainstream definitions of preferences and features extracted from ResearchGate Q&A, four metrics were selected to measure user preferences: question views, follows, answers, and recommendations. To account for temporal effects, the number of months between the question posting date and data collection date was calculated. If the interval was less than one month, it was recorded as 1 month. Each metric (views, follows, answers, recommendations) was divided by the time interval (in months) and then standardized to derive normalized scores. A pairwise correlation analysis of the four metrics revealed low inter-metric correlations ($r < 0.7$, $p < 0.01$). Consequently, the public preference score for each question was defined as the sum of the time-averaged and standardized values of views, follows, answers, and recommendations. Questions were proportionally divided into "high" and "low" preference tiers based on their aggregated scores, resulting in 5,000 popular questions and 5,000 unpopular questions.

The subsequent step involved constructing the fine-tuning dataset. Each data instance comprised three components: a prompt, an input, and an output. The input consisted of the textual content of each question, which was further divided into two configurations for comparative analysis: (1) question title only, and (2) question title combined with its detailed description. This dual-input approach was designed to evaluate the impact of varying contextual information on prediction performance. The output represented the public preference level for the question, categorized as either "high" or "low." A total of 10,000 questions were randomly sampled and split into training and testing sets at an 8:2 ratio (8,000 for training and 2,000 for testing), with equal representation of both preference categories in each subset to ensure class balance.

Finally, three widely adopted and high-performing base models—GPT-4o-mini (OpenAI), DeepSeek-R1-Distill-Qwen-7B (DeepSeek), and Gemini 1.5 Flash (Google)—were selected for experimentation. The training set was used to fine-tune these models, followed by performance testing to assess their predictive capabilities.

Table 2. Performance Evaluation Results of the Models.

Input	Model name	Popularity Level	Acc (%)	F1 (%)	P (%)	R (%)
Title	GPT-4o-mini	high	70.4	71.4	69.3	73.6
		low		69.4	71.8	67.1
		average		70.4	70.5	70.4
	DeepSeek-R1-Distill-Qwen-7B	high	71.0	70.5	71.7	69.3
		low		71.5	70.3	72.7
		average		71.0	71.0	71.0
	Gemini 1.5 Flash	high	67.8	71.7	63.9	81.7
		low		62.6	74.7	53.9
		average		67.2	69.3	67.8
Title + Description	GPT-4o-mini	high	71.7	72.8	69.9	76.0
		low		70.4	73.8	67.3
		average		71.6	71.9	71.7
	DeepSeek-R1-Distill-Qwen-7B	high	72.7	71.8	74.2	69.6
		low		73.5	71.4	75.8
		average		72.7	72.8	72.7
	Gemini 1.5 Flash	high	72.6	73.6	70.9	76.5
		low		71.4	74.5	68.6
		average		72.5	72.7	72.6

Results

The performance evaluation results of the models are summarized in Table 2. These findings indicate that LLMs exhibit promising potential in predicting user preferences for academic questions, achieving an average prediction accuracy of approximately 70%. Specifically, in the task of predicting preferences based solely on question titles, the DeepSeek-R1-Distill-Qwen-7B model

delivered the best performance, with an accuracy of 71%, while Gemini 1.5 Flash showed comparatively weaker results, achieving 67.8% accuracy. When the input context was expanded from titles only to titles + descriptions, all three models exhibited performance improvements. This confirms that providing richer contextual information enhances LLMs' predictive capabilities. Notably, Gemini 1.5 Flash demonstrated the highest improvement, with a 4.8% increase in accuracy. In contrast, DeepSeek-R1-Distill-Qwen-7B showed a more modest gain of approximately 1% when supplemented with descriptive text. These findings suggest that DeepSeek and GPT-4o-mini may excel at processing concise title-based inputs, where additional detailed information from longer question descriptions contributes marginally to accuracy. Conversely, Gemini 1.5 Flash appears better equipped to leverage complex inputs, effectively integrating both titles and descriptions to refine its predictions.

Conclusions

This study investigates the feasibility and accuracy of LLMs in predicting users' academic information preferences based on textual content from questions on ResearchGate Q&A, an academic social Q&A platform. The findings reveal that, compared to traditional machine learning algorithms reliant on feature engineering, LLMs achieve higher accuracy in predicting user preferences (Li et al., 2015; Li et al., 2020), and providing richer textual information (e.g., question descriptions) positively enhances their predictive performance. Among the tested models, DeepSeek-R1-Distill-Qwen-7B delivered the best results under both input conditions (title-only and title+description), while Gemini 1.5 Flash demonstrated the most significant performance improvement (4.8%) when additional detailed context was introduced. In conclusion, this work validates the preliminary efficacy of LLMs in predicting academic information preferences and provides insights for optimizing LLMs-as-judges in diverse application scenarios. Future research could incorporate external question features, such as the objective attributes of question askers (e.g., expertise, institutional affiliation), to enable more precise question popularity prediction. Furthermore, extending the evaluation of LLMs' predictive capabilities to other social media platforms would strengthen the generalizability of these findings.

References

Liao, D., Xu, J., Li, G., Huang, W., Liu, W., & Li, J. (2019, July). Popularity prediction on online articles with deep fusion of temporal process and content features. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 200-207).

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., ... & Liu, Y. (2024). Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Li, L., He, D., Jeng, W., Goodwin, S., & Zhang, C. (2015, May). Answer quality characteristics and prediction on an academic Q&A Site: A case study on ResearchGate. In *Proceedings of the 24th international conference on world wide web* (pp. 1453-1458).

https://doi.org/10.51408/issi2025_171

The Impact of Brazilian Scientific Production on Public Policies: A Scientometric Analysis

Bernardo Cabral[1], Carlos Graziani[2], Evandro Cristofoletti[3], Guilherme Macari[4], Karen Esteves Fernandes Pinto[5], Sergio Salles-Filho[6], Yohanna Juk[7]

[1]*bernardopcabral@gmail.com*
Federal University of Bahia (Brazil)

[2]*carlosgraziani.toledo@gmail.com*
Faculty of Social and Applied Sciences of Extrema (Brazil)

[3] *evcoggo@unicamp.br,* [4]*g217133@dac.unicamp.br,* [5] *karenefp@unicamp.br,* [6]*sallesfi@ige.unicamp.br,* [7]*yjuk@unicamp.br*
State University of Campinas (Brazil)

Introduction

The integration of scientific knowledge into public policies is a critical pathway for addressing complex societal challenges, such as health crises, environmental sustainability, and economic inequality. Understanding how science contributes to policymaking is essential for assessing the societal impact of research and enhancing evidence-based decision-making processes (Bozeman & Youtie, 2017; Dorta-González et al., 2024). However, evaluating the direct influence of scientific outputs on policy remains challenging, particularly in countries of the Global South, where geographic and linguistic biases often limit visibility in international databases.

In Brazil, despite significant challenges in funding and conducting research, the country has developed a robust scientific infrastructure, producing a high volume of publications across diverse disciplines. This study investigates the extent to which Brazilian scientific output is cited in public policy documents globally, leveraging the Web of Science (WoS) to identify relevant publications and the Overton database to track their impact on policies.

Building on methodologies explored in prior studies (Cristofoletti et al., 2024) this research not only quantifies the use of Brazilian science in policymaking but also identifies the thematic areas where its impact is most prominent. By focusing on the intersection of research and policy, the study contributes to discussions on enhancing the societal relevance of science and addressing methodological challenges in impact evaluation.

Method

We employed bibliometric methods to analyze metadata from research articles authored by Brazilian researchers indexed in the Web of Science (WoS). The records were identified using search strategies in the WoS advanced search mode, applying the field tag CU=Country/Region and selecting only articles associated with "Brazil" or "Brasil." This approach ensures that every article included has at least one Brazilian author. The dataset spans from 1900 to 2023 and was retrieved on

The search resulted in a total of 964,075 articles. We extracted the Digital Object Identifiers (DOIs) of all records for further analysis. These records were imported into data analysis tools for organization and treatment.

The data included metadata fields such as publication year, research areas (subject categories assigned by WoS), journals, author affiliations, funding agencies, and countries of co-authors. To ensure consistency and accuracy, we cleaned and standardized the data on authors' affiliations (from here on, organizations) and funding agencies using text-mining software and manual verification.

The DOIs of these publications were then queried in the Overton database, which

indexes public policy documents and their references. Overton's API was used to extract data on mentions of Brazilian publications in policy documents. After retrieval, the data underwent cleaning and verification to ensure consistency and remove duplicates. This process allowed us to link each scientific publication to policy documents that cited it as previously suggested in the literature (Cristofoletti et al., 2024).

Results

This study provides a comprehensive analysis of Brazilian scientific production and its impact on public policy documents. The findings are structured into two subsections: the first explores the profile and characteristics of Brazil's scientific output, while the second examines the extent to which this research is cited and utilized in policymaking contexts globally.

Brazilian scientific production

The temporal analysis of Brazilian scientific production reveals consistent growth over the decades, with an acceleration after 2010. The most productive year was 2021, with 76,884 publications. By 2023, 964,075 publications with at least one Brazilian author were indexed in the Web of Science. In terms of international collaboration, the United States stands out as the most frequent partner, with 111,640 publications, followed by the United Kingdom (43,608), France, Germany, and Spain.

Institutions play a central role in Brazil's research output. The University of São Paulo (USP) leads with 203,238 publications, followed by other major public universities such as the State University of Campinas (UNICAMP) (69,305), the Federal University of Rio de Janeiro (UFRJ) (62,743), and Sao Paulo State University (UNESP) (62,201). Additionally, organizations like the Fundação Oswaldo Cruz (FIOCRUZ) and the Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) contribute significantly to areas such as health and agriculture.

Research funding is heavily supported by national agencies, particularly the National Council for Scientific and Technological Development (CNPq), which funded 297,840 publications, and the Coordination for the Improvement of Higher Education Personnel (CAPES), which supported 198,333. State-level agencies like the São Paulo Research Foundation (FAPESP) (124,065) and international organizations, such as the National Science Foundation (NSF) (12,998) and National Institutes of Health (NIH) (10,931), also play significant roles. This comprehensive support network highlights the foundation for Brazilian science and its integration into global research initiatives.

Policy impact

The number of policy documents citing articles with at least one researcher affiliated with a Brazilian institution has grown significantly over time, totaling 161,693 documents. While citations were relatively low before 2010, a steady increase is observed, peaking in 2021 with 18,873 documents.

The majority of policy documents citing Brazilian research were written in English (71.8%), followed by Portuguese (9.2%) and Spanish (7.9%). In terms of authorship, governments were responsible for 41% of these citations, while international governmental organizations (IGOs) accounted for 32.9%, and think tanks contributed 18.8%. Figure 1 highlights that intergovernmental organizations (IGO) and major countries like the USA and the UK are among the most frequent citers, with Brazil itself ranking third.

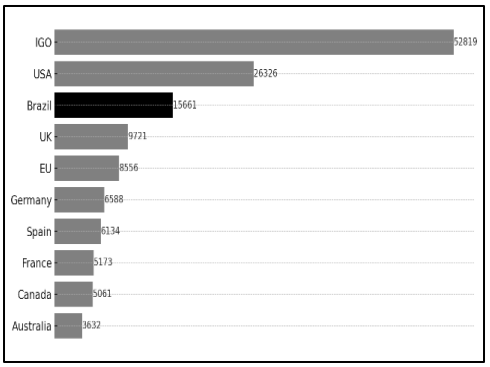


Figure 1. Citations of Brazilian Research in Public Policy Documents by Country.

Figure 2 shows that global institutions such as the World Health Organization (WHO), Food and Agriculture Organization of the United Nations (FAO), and the World Bank are the ones that reference Brazilian research the

most. Brazilian institutions like *Instituto de Pesquisa Econômica Aplicada* (IPEA) and *Fundação Getulio Vargas* (FGV) also contribute substantially, alongside the Brazilian Government.

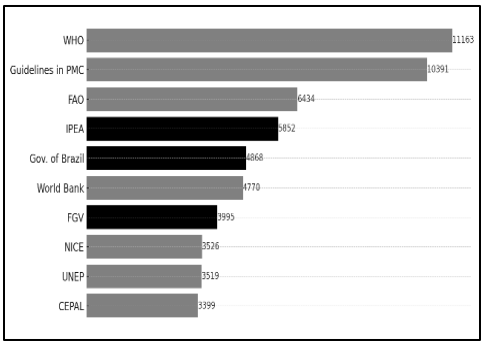


Figure 2. Institutions That Most Cited Brazilian Research in Public Policy Documents.

Conclusion

This study highlights the growing influence of Brazilian scientific research on global policymaking, demonstrating a steady increase in citations within policy documents over time. The findings reveal that Brazilian research is widely referenced, particularly in English-language documents and by international organizations, governments, and think tanks. While major global institutions such as the WHO and World Bank play a significant role in citing Brazilian research, national institutions like IPEA and FGV and the Brazilian Governemnt itself also use Brazilian-made science. These results show how Brazilian science is used for policies, despite challenges in research funding and visibility. Strengthening mechanisms to enhance the accessibility and influence of Brazilian research can further expand its role in global policy debates.

Acknowledgments

This research is funded by the São Paulo Research Foundation (FAPESP).

References

Bozeman, B., & Youtie, J. (2017). Socio-economic impacts and public value of government-funded research: Lessons from four US National Science Foundation initiatives. *Research Policy*, *46*(8), 1387–1398. https://doi.org/10.1016/j.respol.2017.06.003

Cristofoletti, E. C., Salles-Filho, S., Juk, Y., Cabral, B., Pinto, K. E. F., Hollanda, S., Graziani, C., & Pereira, C. A. (2024). A long and winding road: Research impact evaluation over public policies. *Quantitative Science Studies*, 1–28. https://doi.org/10.1162/qss_a_00345

Dorta-González, P., Rodríguez-Caro, A., & Dorta-González, M. I. (2024). Societal and scientific impact of policy research: A large-scale empirical study of some explanatory factors using Altmetric and Overton. *Journal of Informetrics*, *18*(3), 101530. https://doi.org/10.1016/j.joi.2024.101530

https://doi.org/10.51408/issi2025_194

Trends and Distribution of Domestic and International Research Collaboration: An Asian View

Szu-Chia S. Lo[1], Mu-Hsuan Huang[2]

[1] *szuchialo@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd. 106319 Taipei (Taiwan)

[2] *mhhuang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd. 106319 Taipei (Taiwan)

Introduction

It has been decades that collaboration is taken as a good strategy in research development, and attracts the researchers or policy makers to either take actions to exercise the collaborative strategies or set-up the criteria to encourage the action (Katz & Martin, 1997; Kyvik & Reymert, 2017; Ponomariv & Boardman, 2016). In this study, the authors saw the co-authorship as a presentation of research collaboration and took the universities from Asian countries as tokens to investigate the following themes from a macro view, country, and observe the impact of COVID-19 on the collaboration. The following are the research questions targeted in this work.

-The trends of the collaborations in research of Asian countries.

-The distributions of the domestic and cross-countries collaborations in research of Asian countries

-The similarity of the collaborative actions among the universities that are with different research strength of Asian countries

Method and data

The authors adopted the bibliometrics approach, and the description of the details of the research design is followed.

Study informants

In this study, the authors obtained a list of Asian universities from the NTU World University Rankings and gained the lists of universities from 30 Asian regions as the study targets, and the publications done by the affiliated members of the selected universities included in this study published in 2017 and 2022 were searched from WoS for the further analysis.

Indicators and data process

The following indicators, such as CCj and CoCj, were developed to present the results statistically.

-*Research Productivity Index*. **CCj**, Number of publications count by country j (CCj)

$CCj = \sum_{i=1}^{n} Pi$, i=1to n, n=number of universities of the target country. P*i*=publication count of the university *i of country j.*

-*Collaborative Effort Index (Inter-institutional co-authored publications identifying and tagging)*, **CoCj**, Number of co-authored of publications count by country j (CoCj)

$CoCj = \sum_{i=1}^{n} CoUi$, i=1to n, n=number of universities of the target country, CoU*i*=co-authored publication counts of the university *i of country j.*

-*Domestic and international collaboration detecting and tagging*

The information of all the affiliations of the co-authors was paired accordingly, and if the affiliations are the same or located in the same region, the work would be marked domestic collaborative works. The work would be marked as a cross-country collaborative effort If different country information showed.

Results

The following are the preliminary results of the analysis, which includes the productivity, collaboration, and the distribution of the collaboration by domestic and international. The discussion starts from an overall viewpoint and further breaks down to three productive tiers, and all from a macro level-countries point of view.

Affiliations and productivities

There were around 2 million (2017, 726,323; 2022, 1,276,611) scholarly publications that were done by the associates of the designated universities.

Single and collaborative research effort

High percentage research collaboration. Figure 1 shows the distribution of the single authored and co-authored papers by the presentation of the percentages, including total count, average, minimum and maximu m results are shown.

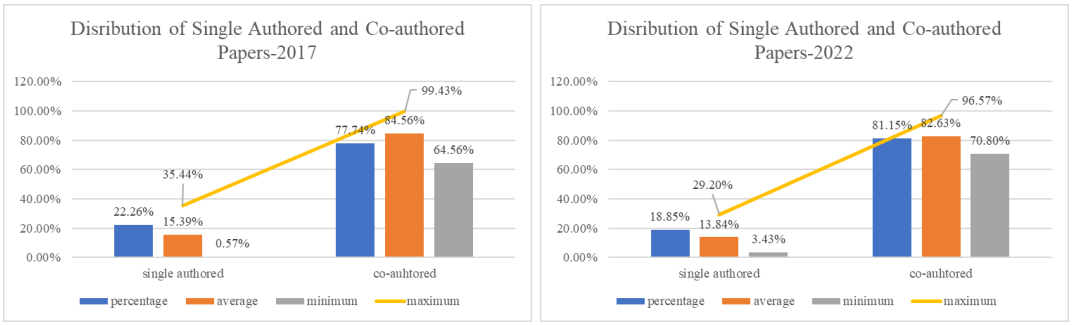


Figure 1. Distribution of Single Authored and Co-authored papers, 2017 and 2022.

Domestic and international collaboration

Domestic collaboration was preferred, but the further analysis indicates the diverse strategies in the research collaboration from region to region. The distribution of the percentages of the domestic authored works of the collaborative works was from 2.87% to 63.25% in 2017 and 3.05% to 69.11% in 2022, and the one for cross-countries collaboration was from 36.75% to 97.13% in 2017 and 30.89% to over 96.95% in 2022. (Figure 2)

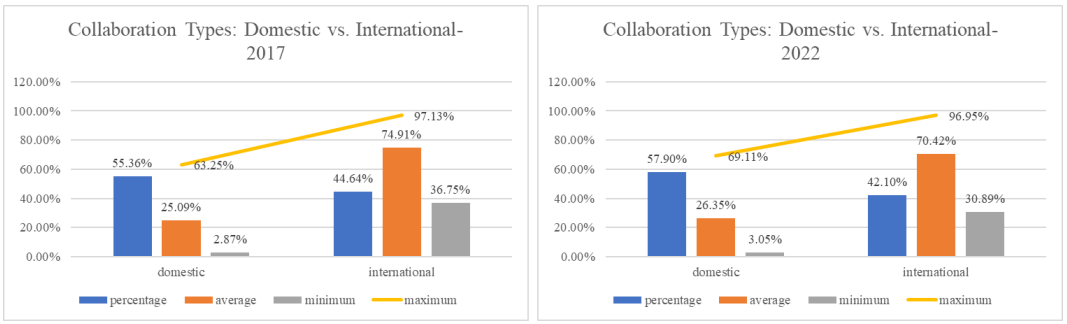


Figure 2. Distribution of Collaboration types: Domestic vs. International Collaboration, 2017 and 2022.

Collaboration and productivity

The study further investigated if the different collaborative strategies might be taken by various t research productive entities, the authors grouped the regions into three tiers, core, intermediate, and peripheral, by the research productivities, by referencing the idea of Bradford Law. The results show that the universities of the three different tiers all have more research work done by collaborating with peers from the academic community. It was found that several areas do either attract or rely on the knowledge or resources input from other areas in the research work. (Figure 3)

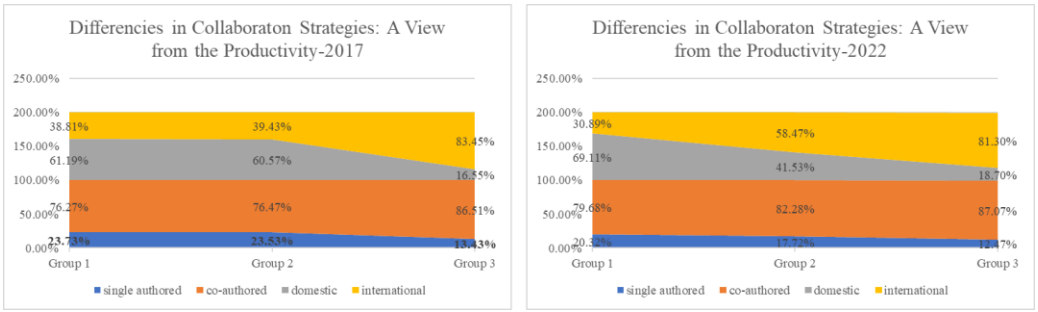


Figure 3. Collaboration Preferences: A View from the Different Productivities.

Conclusions and discussion

The results of this study show the high percentage of research outputs were done under the collaborative efforts, the collaborative scenes were not changed under the impact of COVID-19, for the two sampled years. Generally speaking, domestic collaboration is preferred, however there is evidence that the universities with less research productivity do devote more effort into international collaboration.

Acknowledgments

The Center for Research in Econometric Theory and Applications, the Featured Areas Research Center Program, Ministry of Education (MOE) in Taiwan.

References

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*, 1-18. https://doi.org/10.1016/S0048-7333(96)00917-1.

Kyvik, S., & Reymert, I. (2017). Research collaboration in groups and networks: Differences across academic fields. *Scientometrics, 113*, 951-967. https://doi.org/10.1007/s11192-017-2497-5

Ponomariov, B, & Boardman, C. (2016). What is co-authorship? *Scientometrics, 109*, 1939-1963. https://doi.org/10.1007/s11192-016-2127-7

https://doi.org/10.51408/issi2025_184

Unraveling Evolutionary Dynamics of Disruptive Innovations: Insights from Multi-Scale Knowledge Networks

Haiyun Xu[1], Junhao Yang[2], Xiao Tan[3], Shuying Li[4], Zenghui Yue[5], GuotingYuan[6], Xin Li[7]

[1] *xuhaiyunnemo@gmail.com*, [2]*yangjunhao163@gmail.com*, [7]*yian1207@163.com*
Business School, Shandong University of Technology, Zibo, 255000 (China)

[3]*tantan46227@163.com*
Beijing Institute of Science and Technology Information, Beijing Academy of Science and Technology, Beijing, 100089 (China)

[4]*lisy@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[5]*yzh66123@126.com*
School of Medical Information Engineering, Jining Medical University, Rizhao (China)

[6]*guotingyuan@hotmail.com*
School of Foreign Languages, Jining Medical University (China)

Introduction

Identifying early features of disruptive innovations is crucial for shaping science and technology strategies but remains difficult due to their interdisciplinary nature and weak early signals (Ioannidis, Cristea, & Boyack, 2020). Current methods often rely on retrospective metrics and lack predictive power to guide innovation planning. This study uses time-series network analysis to investigate the micro- and meso-scale evolution of knowledge structures. By revealing sequential patterns and cross-scale linkages, it uncovers the mechanisms driving the early emergence of disruptive innovations, enhancing foresight into their formation and development.

Related research and research gaps

Time-series networks reveal scientific knowledge evolution and disruptive innovation patterns through community dynamics and weak signal detection (Ceria et al., 2022). Multi-layer and motif-based analyses expose interactions across micro- and meso-levels. Yet, research often overlooks cross-scale transitions and multi-level dynamics (Lobbé et al., 2022). This study addresses these gaps by analyzing multi-scale knowledge evolution to identify early drivers of disruptive innovations.

Materials and methods

This study analyzes the evolution of disruptive innovation themes by examining multi-scale knowledge types at micro (nodes, edges, motifs) and meso (thematic communities) levels. Using time-sliced dynamic networks and multi-layer temporal analysis across basic, applied, and industrial research, it constructs knowledge evolution sequences. Techniques like dynamic time warping and anomaly detection identify consistent patterns and mutations in knowledge sequences, revealing the driving mechanisms behind the early emergence and development of disruptive innovation within evolving knowledge networks.

Data Collection and Processing

This study uses scientific papers, patent literature, and product information as carriers to represent scientific, technological, and industrial knowledge content, respectively. At the same time, abstract data is used as the specific data object for constructing the knowledge network. Subsequently, natural language processing techniques are applied to

structure the collected text data and extract thematic keywords with entity-concept significance.

Construction of Temporal Knowledge Networks and Knowledge Type Sequences

This study constructs temporal knowledge networks by creating time slices and building thematic keyword co-occurrence networks representing basic research, applied research, and industrial application stages of disruptive innovation. It quantifies nodes, edges, motifs, and community structures within each time slice, using the Fast GN algorithm for community detection and the BEAM method with Orca for motif identification. These analyses generate multivariate time series of 32 knowledge structure types, enabling the study of multi-scale knowledge type evolution sequences and their interactive patterns.

Results and Discussion

This study selects the field of regenerative medicine (stem cells) as the empirical domain for early weak signal identification in the process of structural changes in emerging topic networks. Stem cells are an emerging field of medicine that aims to develop technologies capable of regenerating, repairing, or replacing damaged (diseased) cells, organs, tissues, etc. The current field of leukemia treatment mainly focuses on incremental improvements along existing technological trajectories, exhibiting distinct progressive technological characteristics. This field is selected as the control domain.

Knowledge Type Sequence Evolution Consistency Analysis Based on DTW

This study applies Dynamic Time Warping (DTW) (H. Li et al., 2020) to single-scale knowledge type sequences from scientific, technological, and industrial knowledge networks in SC and LK fields. After z-score normalization, DTW distances are calculated to assess sequence consistency. The resulting DTW distance matrix is visualized as a heatmap (Figure 1), where lighter colors indicate higher similarity and darker colors indicate lower similarity.

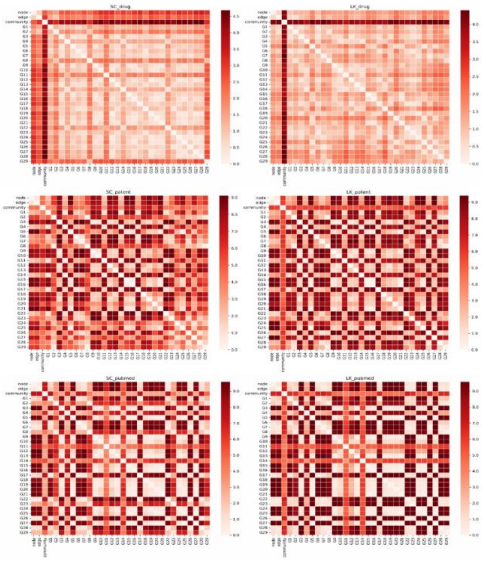


Figure 1. DTW Distance Matrix Heatmap of Knowledge Type Sequences in the Industry-Technological-Scientific Knowledge Network.

Multiscale Knowledge Type Sequence Evolution Consistency Analysis

DTW analysis reveals that scientific and technological knowledge type sequences evolve more consistently within each domain, with the smallest DTW distances observed between these two networks. The SC domain shows slightly higher consistency in technological and industrial knowledge evolution compared to LK. Across both domains, industrial knowledge types exhibit the greatest overall consistency, followed by technological, while scientific knowledge types show the lowest consistency. However, when comparing inconsistency levels, the industrial sector displays the most variation, highlighting distinct evolutionary dynamics and driving mechanisms across science, technology, and industry (Figure 2).

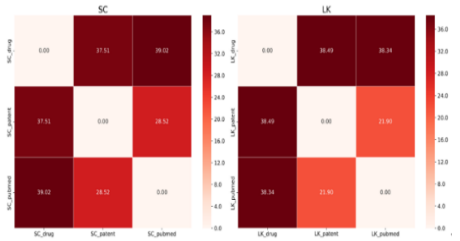


Figure 2. Heatmap of DTW Distances Between Multiscale Knowledge Type Sequences.

Conclusion and policy implications

This study investigates the multi-scale co-evolution of knowledge types in early-stage thematic networks to reveal how disruptive innovation themes emerge and develop. Using temporal network analysis, it constructs micro- and meso-scale knowledge sequences across basic, applied, and industrial research stages. Dynamic Time Warping (DTW) and anomaly detection methods quantify evolutionary consistency and identify mutation timings, showing higher consistency in industrial networks and domain-specific differences between stem cell and leukemia fields.

Acknowledgments

This article is supported by the National Natural Science Foundation of China (No.72274113), Shandong Provincial Natural Science Foundation (No. ZR202111130115), Shandong Provincial Social Science Foundation (No.23CTQJ07), Beijing Natural Science Foundation (No.9242006) and the Taishan Scholar Foundation of Shandong province of China (tsqn202103069).

References

Ceria, A., Havlin, S., Hanjalic, A., & Wang, H. (2022). Topological–temporal properties of evolving networks. *Journal of Complex Networks*, 10(5), cnac041.

Ioannidis, J. P., Cristea, I.-A., & Boyack, K. W. (2020). Work honored by Nobel prizes clusters heavily in a few scientific fields. *Plos One*, 15(7), e0234612.

Li, H., Liu, J., Yang, Z., Liu, R. W., Wu, K., & Wan, Y. (2020). Adaptively constrained dynamic time warping for time series classification and clustering. *Information Sciences*, 534, 97–116.

Lobbé, Q., Delanoë, A., & Chavalarias, D. (2022). Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge. *Information Visualization*, 21(1), 17–37.

Author Index

Achal Agrawal 1945
Adel Diyaf 1375
Aditi Ashok 2050
Adolfo Alonso-Arroyo 2409, 2448, 2451
Ai Kishimoto 2291
Akshat Nagori 1337
Alena Nefedova 2169
Alesia Zuccala 2035
Alexander Dmitrienko 2130
Alexander J. Gates 1745, 2058
Alexander Karpov 2511
Alexander Libman 2299
Alexander Schniedermann 2199
Alexandra Malysheva 2185
Alexandre Clausse 2367
Alexey Zheleznov 2237, 2514
Andrea Sixto-Costoya 2409
Andrey Guskov 2185, 2406
Andrey Lovakov 1736, 2199
Andrey Zayakin 2299
Angelika O. Tsivinskaya 2154
Anna Abalkina 2299
Annalisa Di Benedetto 1593
Antonio Malo 2374
Antonio Zinilli 1717
Aram Mirzoyan 1535, 2397
Artur Pecherskikh 2154
Ayush Tripathi 1945
Balázs Győrffy 1841
Barbara S. Lancho Barrantes 2466
Beate Rusch 2439
Beibei Sun 2359
Bernardo Cabral 2259, 2520
Boglárka Weltz 1841
Carlos Areia 1972
Carlos Graziani 2520
Carolina Coimbra Vieira 1995
Cassidy R. Sugimoto 2050, 2138, 2193
Catarina Carreira 2484
Ceren Bilge Seferoğlu 2267
Changcheng Xue 2400
Chaocheng He 2103, 2306
Chao-Chih Hsueh 1361
Chen Shu 2502
Chen Yunwei 1660, 1978, 2475, 2505
Chenchen Huang 1440
Chengzhi Zhang 1813, 1899, 2457, 2095
Chiaki Miura 2073
Chien Hsiang Liao 2469
Christoph Schindler 2481
Christophe Malaterre 2275
Chwen-Li Chang 1952
Cinzia Daraio 2374, 2454, 2460, 2508
Ciriaco Andrea D'Angelo 1717
Cristian Mejia 1471
Cristiana Agapito 2484
Cristina Arhiliuc 1699
Cristina Rius 2448, 2451
Dag W. Aksnes 2066, 2146
Dai Xinran 1613

Daniel Karabekyan 2511
Dar-Zen Chen 1361
Dejan Ravšelj 2010
Denis Kosyakov 2185, 2406
Didier Torny 2050
Dimity Stephen 2199
Dingkang Lin 2177
Dmitry Kochetkov 1523, 2406
Edoardo Fazzini 2228
Ekaterina Dyachenko 2237, 2514
Elena Chechik 1995, 2154
Eleonora Dagienė 2121
Elizaveta Chefanova 2169
Elizaveta V. Sokolova 2442
Emanuel Kulczycki 2050, 2079, 2267
Emanuela Reale 1717
Eugenio Oropallo 2454, 2508
Eugenio Petrovich 2228
Evandro Cristofoletti 2259, 2520
Ezgi Uğurlu 2267
Fabio Nonino 2454, 2508
Feicheng Ma 1301
Fidan Badalova 2367
Francis Lareau 2275
Fredrik N. Piro 2066, 2146
Fuzhen Liu 2306
Gabriel Falcini 2490
Gang Li 2387
Gevorg Kesoyan 1535, 2430
Giovanni Abramo 1717
Giulio Maspero 2374
Guilherme Macari 2520
Guillaume Cabanac 2367
Guiyan Ou 2103, 2306
Güleda Doğan 2267
Gunnar Sivertsen 1507
Guo Chen 1987, 2112, 2325, 2496
GuotingYuan 2526
Haiyan Hou 2035
Haiyun Xu 1389, 2526
Haochuan Cui 2002
Haoyu Li 2445
Helan Wu 2478
Henrik Karlstrøm 2066, 2146
Honami Numajiri 2043
Hongrui Yang 2478
Hongyu Wang 2400, 2418
Hu Wei 1415
Hui Fu 1674
Hui Peng 2493, 2517
Hui Zhang 1674
Ichiro Sakata 2073
Ida Svege 2146
İdris Semih Kaya 2267
Ilaria Vigorelli 2374
Ilya Gorelskiy 2511
Indraneel Mane 1745
Irina Lakizo 2406
Irina Selivanova 2185
István Szabó 1841
Jack H. Culbert 2283, 2415
Janina Zittel 2439
Jean-Charles Lamirel 2275
Jesús María Godoy 2251
Jiahao Li 1440, 1389, 2177
Jiang Wu 2103, 2306
Jianhua Hou 2193
Jianjian Gao 1745, 2058
Jiaqi Zeng 2095
Jiaxing Li 1849
Jiayi Hao 1813
Jin Mao 2387
Jin Xiaohong 2502
Jingyuan Li 2095
Jinyu Gao Yi Bu 2087

Jinzhu Zhang 2445
Johann Mouton 2050
Jorge Gulín-González 1978
Juan Rogers 2050
Judit Hermán 2433
Julián D. Cortés 2251
Junhao Yang 2526
Junwan Liu 1440
Kai Li 2002
Kaile Gong 2487
Kaiwen Shi 2400, 2418
Karen Esteves Fernandes Pinto 2259, 2520
Katerina Guba 2154, 2237, 2514
Kathryn O. Weber-Boer 1972, 2244
Kieron Flanagan 2050
Kíra Diána Kovács 2433
Kiran Sharma 1337, 1482, 1582, 2472
Kuei Kuei Lai 1952
Lan Umek 2010
Lei Li 2493, 2517
Lele Kang 1849
Li Hanxi 1323, 2421
Li Hui 2502
Li Jiake 1770
Li Jian 1770, 1794
Li Jiangbo 1770, 1794
Li Jiayu 2502
Li Li 1493, 2207
Li Xinran 1415
Liang Shuang 1323, 2421
Lili Wang 1964
Lin Zhang 1507
Linlei Xie 1899
Liu Hao 1660
Liu Xiaojuan 1415, 1613
Liu Xiaoping 2421
Liu Yajing 1323
Liyue Chen 2161
Lorenzo Gandolfi 2228
Lu DONG 2214
Luiza Petrosyan 2409
Manya 1337
Maria Ohanyan 2397, 2430
Maria Yudkevich 2130
Mariam Yeghikyan 2397, 2430
Marion Schmidt 2199
Mathieu Ouimet 2050
Matteo Ottaviani 2199
Maxim Dmitriev 2237, 2514
Mehul Dubey 1337
Michio Oguro 2043
Mingxia Lu 2445
Mingze Zhang 1555, 1964, 2214
Miranush Kesoyan 1535, 2397, 2430
Moumita Koley 1945
Mu Yingyu 1770
Mu-Hsuan Huang 2523
Nataliya Matveeva 1636, 2130
Nikita Buravoy 2154
Nikita Sorgatz 2199
Niu Shihang 1794
Noor Jaleel 2050
Orsolya Vásárhelyi 2433
Ouyang Wenhao 1794
Özge Söylemez 2267
Parul Khurana 1337, 1482, 1582, 2472
Pei-Shan Chi 2403
Pei-Ying Chen 2050, 2138
Penghui LYU 1555
Peter Aspeslagh 2333, 2427
Philipp Mayr 2283, 2367, 2415

Philippe GORRY 2424
Przemysław Korytkowski 2079
Qingshan Zhou 2436
Qiqi Zhang 1440
Raf Guns 1699
Raf Guns 2478
Rafael Aleixandre-Benavent 2409, 2448, 2451
Reem Abusanina 1375
Roberto Cruz Romero 2199, 2340, 2349
Ronald Rousseau 2463, 2478
Ruinan Li 2359
Rut Lucas-Domínguez 2409, 2448, 2451
Ruzanna Shushanyan 2430
Sanfa Cai 2478
Sarah Bratt 2087
Sergio Luiz Monteiro Salles Filho 2050, 2490
Sergio Salles-Filho 2259, 2520
Shen Jianing 1613
Shi Chen 2493
Shiqi Tang 2193
Shuang Liang 2436
Shuo Xu 1440
Shuya Chen 2112
Shuying Li 1389, 2526
Simon Hunanyan 1535, 2397
Simone Di Leo 2454, 2460, 2508
Siqi Hong 2496
Sisi Li 2478
Sitong Xiang 2306
Stephan Stahlschmidt 2199, 2349
Stephen Wu 1375
Szu-chia Lo 2499, 2523
Takayuki Hayashi 2043, 2291
Tamara Heck 2481
Tamarinde Haven 1972
Tan Fu 1925
Tan Xiao 2502
Tao Zhiyu 1323, 2421
Thorsten Koch 2439
Tim C. E. Engels 1699, 2333
Tindaro Cicero 1593
Tingcan Ma 2359
Tomasz Stompor 2439
Torger Möller 2199
Tzu-Kun Hisao 2138
Valeria Aman 2199
Verena Weimer 2481
Victor A. Blaginin 2442
Victoria Di Césare 1995
Viktor Glukhov 2406
Vladimir Batagelj 1636
Wang Kaile 2475, 2505
Wen Lou 1925, 2412
Wenting Tang 2412
Wolfgang Glänzel 2403
Xi Chen 2387
Xi Guiquan 2502
Xian Zhang 1389
Xianjiang Deng 2193
Xiao Tan 2526
Xiao Yuntong 1415
Xiaofei Li 1987
Xiaoguang Wang 2400, 2418
Xiaojun Hu 2463
Xiaoli Chen 1870
Xiaomin Liu 2161
Xiaoyun Gong 1440
Xin Li 2526
Xin Zhang 1389
Xinyue Lu 1493
Xu Haiyun 2502
Xuehua Wu 2387

Xuezhao Wang 1870
Yafang Fan 2478
Yajie Wang 2035, 2251, 2433
Ye Zhang 2306
Yi Xiang 2457
Yi Zhao 1899, 2095
Yifan Yang 2325
Yiming Liu 2448, 2451
Ying Huang 1507, 1674
Yizhan LI 1555, 1964, 2214
Yoed N. Kenett 2283
Yohanna Juk 2259, 2490, 2520
Yu Liao 2207
Yu Yao 1613
Yuan Sun 2499
Yuanxun Li 2418
Yuanyuan Shang 1507
Yu-Chun Hsu 1952
Yue Hu 2493, 2517
Yue Li 1849
Yuxian Liu 2478
Yuzhuo Wang 2002, 2359
Zehra Taşkın 2267
Zenghui Yue 2526
Zexia LI 1555, 1964, 2214
Zhang Biao 1660
Zhang Mingyue 1794
Zhang Ting 2502
Zhe Cao 1507
Zhesi Shen 1493, 2207
Zhiyu Tao 2436
Zhou Haichen 1978
Ziya Uddin 2472
Zizuo Cheng 1301