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.

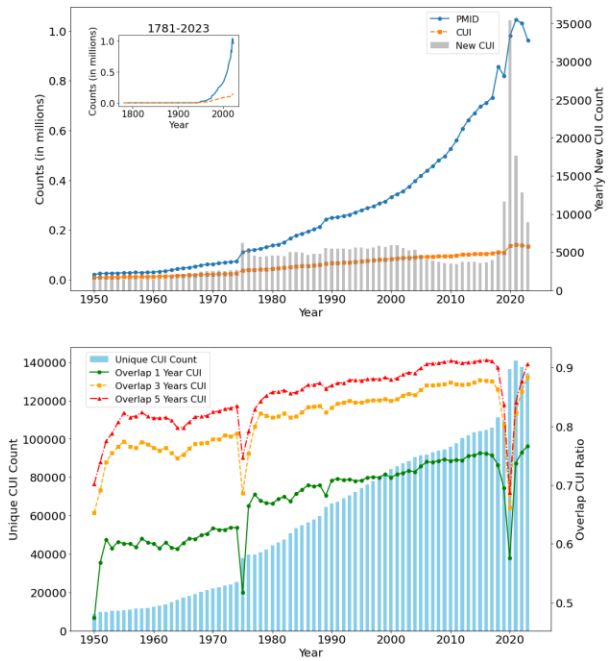

Figure 1. Summary of PMID, CUIs, and Yearly New CUIs in SemMedDB (1950-2023).

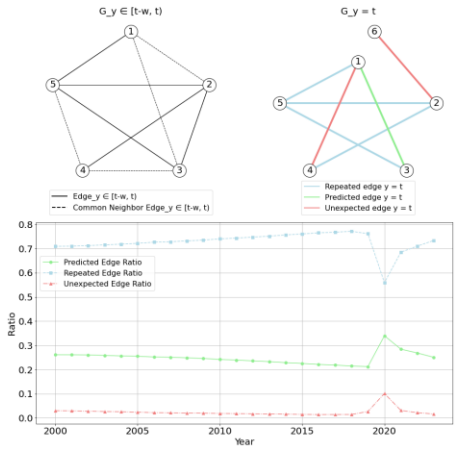

Figure 2. Edge types in concept graph: repeated, predicted, and unexpected.

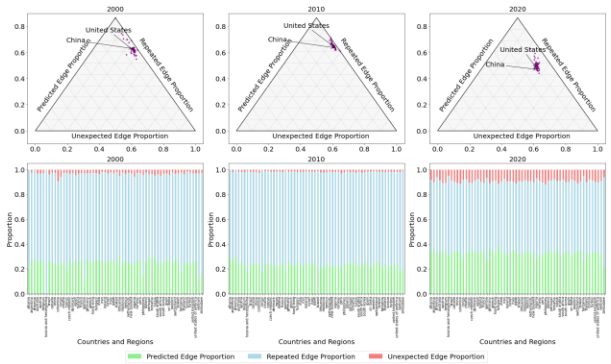

Figure 3. Distribution of Edge Types among Countries.

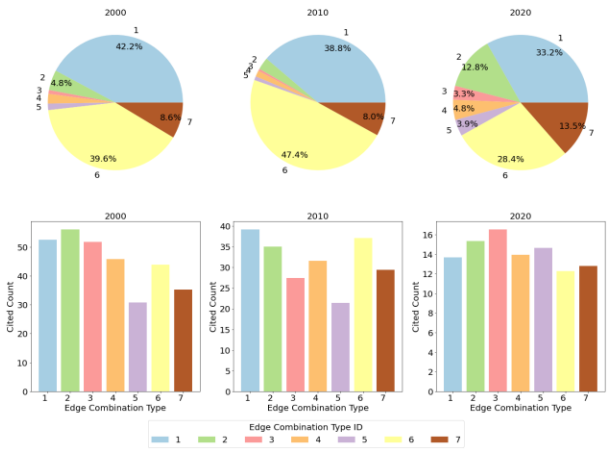

Figure 4. Edge Combination Types.

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