https://doi.org/10.51408/issi2025_096

Multi-Disciplinal, Large Scale Mentorship Dataset and Demographics

Chiaki Miura [1], Yoshiyasu Watanabe[2], Tsubasa Sakammoto[3], Hiroshi Hashizume[4]

[1]1t.miura@gnt.place,[2]watanabe.yoshiyasu.2m@kyoto-u.ac.jp,
[3]sakamoto.tsubasa.5n@kyoto-u.ac.jp,[4]hashizume.hiroshi.8z@kyoto-u.ac.jp
Department of Engineering, The University of Tokyo, Bunkyo, Tokyo (Japan)
Office of Research Acceleration, Kyoto University, Kyoto (Japan)

Abstract

Academic genealogy depicts a relationship network of mentor-trainees, which embraces the rich history of knowledge flow through discipline. We matched over 800 thousand researchers to the world's largest academic genealogy database with the open/libre bibliographic database of over 200 million research works through author names, works, and institutions. This allows scientometricians and higher education strategists to conduct comprehensive analyses of researcher mobility, training, institutional bias, and success. The paper also provides the complete descriptive statistics and propensity of the Academic Family Tree dataset.

Introduction

Mentoring in academia is not only an act of learning, but a profound mecha nism of knowledge transmission. As Zuckerman (1977)[9] and others have ob served, academic disciplines are mediated by formal and informal norms, many of which are implicitly transmitted through interactions between young scien tists and their mentors. Because such interactions are important moments of tacit knowledge exchange across academic fields, the genealogy of mentors and trainees provides a quantitative framework for exploring these relationships and their broader impact on academic ecosystems (7; 4; 1) .

Existing research has emphasized the importance of mentorship in academic career development. Studies from various fields have shown that mentors with high mentorship fecundity, who produce many trainees, increase their scien tific legacy through the success of their students. For example, Sugimoto et al. (2011) (8) demonstrated that the field of expertise of a supervisor directly affects the interdisciplinary nature of a student's dissertation, emphasizing the role of mentorship in the formation of intellectual paradigms. Tol (2024) [8], who recently used the Academic Family Tree dataset to integrate the academic lineage of Nobel Prize winners, also points out that the lineage of academic men tors not only promotes excellence, but also leads to close intellectual networks. These insights highlight the depth of the structural impact of mentorship in cultivating groups of elite scientists.

The "Academic Family Tree", pioneered by David and Hayden 2012 (3) for its antecedent known as"Neurotree", provides a unique opportunity to analyze the relationship between mentors and their trainees on an unprecedented scale. This dataset contains bibliographic record of over 876 thousand scientists and 1.8 million

liaisons, and it is possible to understand how mentoring relationships affect scientific productivity and success. Unlike traditional case studies with limited generalizability, this large-scale dataset enables rigorous statistical analysis across disciplines and institutions.

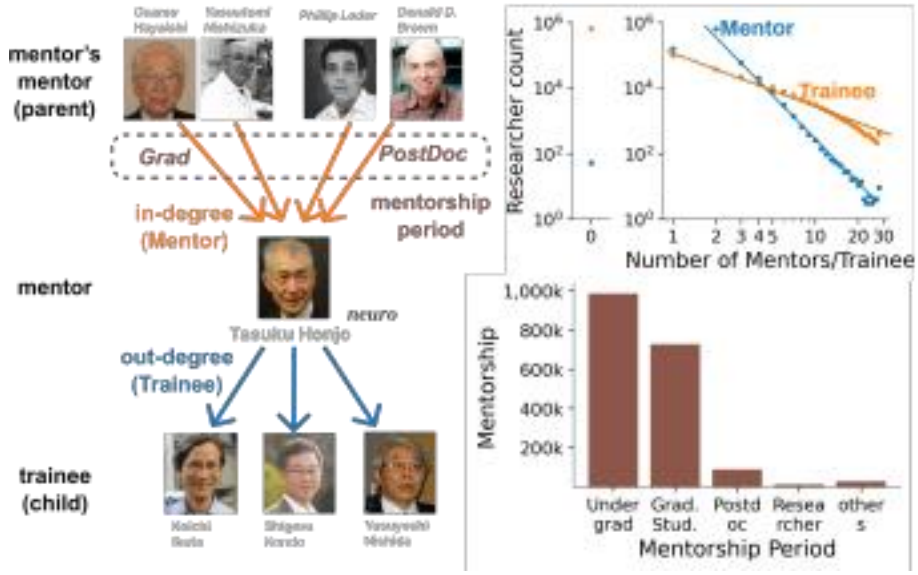

Figure 1: Descriptive statistics of academic family tree dataset.

Ke et al. (2022) (6) combined datasets about mentorship with the Microsoft Academic Graph (MAG) to identify patterns of mentor effectiveness and demographic differences. Building on the findings of these previous attempts, this paper proposes to integrate the Academic Family Tree and OpenAlex – a fully open bibliographic database developed as the successor to the MAG – to present a systematically managed database that allows more scalable analysis of academic genealogy. (2)

Academic family tree is the world's largest human-annotated academic genealogy database. It is later expanded largely by the American dissertation repository (ProQuest). Most of the mentorship relationships registered are mentorships during undergraduate education and graduate student training. It is remarkable considering that the number of graduate students is almost the same as that of postdocs in the same cohort (16,000 in 2000 to 13,000 in three years after that, four major areas aggregated) (5).

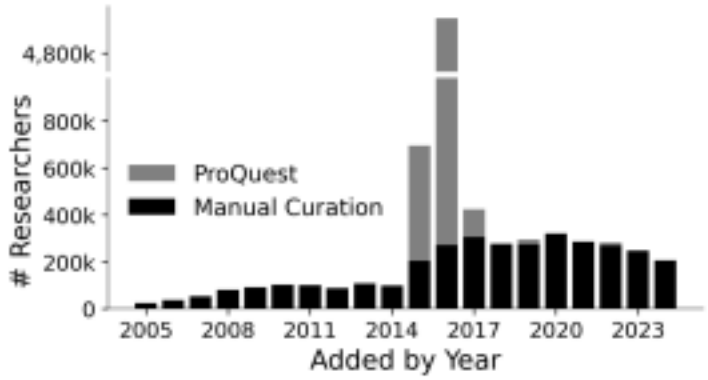

Figure 2. Added by Year.

One of the main questions underlying this research is as follows: What characteristics of the mentor-trainee relationship predict the academic success of the trainee? While previous research suggests that successful mentor is most likely train a successful trainee, the underlying mechanisms are still unclear. Is success primarily a function of intellectual compatibility when the mentor's and trainee's areas of study coincide? Is success due to the mentor's ability to secure access to influential networks and resources? What are the specific mechanisms by which tacit knowledge, such as awareness of grant opportunities or potential collaborators, is transferred from mentor to trainee? Or can these pathways give rise to biases, and how can identifying them help overcome existing barriers to equitable academic advancement? This paper will provide a solid foundation for a more nuanced understanding of the impact of mentorship on academic careers and guide the development of policies and initiatives to support the next generation of scholars.

Table 1. Researcher entity and coverage of Academic Family Tree Attribute Data Count.

Attribute	Data Count	Coverage (%)
First name	875,162	99.87
Middle name	448,272	51.15
Last name	874,974	99.84
Degrees	470,156	53.65
Location	876,147	99.98
Major Area	876,230	99.99
Areas	626,569	71.50
Award	866	0.10
h-index	379,356	43.29
ORCID ID	6,821	0.10
Semantic Scholar ID	379,356	43.28
Homepage	67,519	7.70
Added by	876,228	99.99
Date Added	876,330	100.00

Table 2. Relationship entity and coverage of Academic Family Tree Attribute Data Count Coverage (%).

Attribute	Data Count	Coverage (%)
Mentorship Period	1,841,686	100.00
Location[1]	1,841,103	99.97
Dissertation Title	520,870	28.28
ProQuest ID	438,576	23.81
Added by	1,840,868	99.96
Date Added	1,841,686	100.00

*Location is complemented by Location ID (locid)

OpenAlex

This study maps the authors in the two bibliographic databases and investigates the statistics and registration bias in the AFT dataset. Main source of OpenAlex authors are from the authorship in the works that are mainly retrieved from Crossref, and information about authors "comes from MAG, Crossref, PubMed, ORCID, and publisher websites." OpenAlex then disambiguates and aggregates author records based on how well authors with the same name share a tendency of their works. This algorithm allows us to incrementally aggregate differently written author names and is robust against spelling inconsistencies.

Method and Materials

We used the snapshot of Academic Family Tree (AFT) taken on Oct. 2024. Out of 876,304 researchers on the AFT dataset, 1,168 (0.13%) and 1,356 (0.15%) are missing first name and last name, respectively. We removed records whose first name and last name are both missing, which is equivalent to 920 researchers. We did a few more cleansing, and name and ID normalizations were done to get the best matching accuracy (see Supplementary 1). The whole procedure is depicted in Fig.3. Here, we took ORCID ID as a gold standard, which yields 6,766 matched researchers between the two databases (see supplementary 2). Among the rest, around half (51.2%) have a middle name. Coverage of other major columns are 100%, 98.2%, 53.6%, and 7.3% for major area, location, degree, and homepage, respectively [Table 1.].

As a preliminary result, here we propose a result from the sample of 10,000 AFT records. We conducted all the matching through OpneAlex API between Jan. 5. 2025 and Jan. 10. 2025. In the final version of this matching is done on OpenAlex full snapshot. We first took each author record in AFT dataset, and searched via OpenAlex API using the author's full name as a query.

Figure 3. Matching procedures.

Result and Discussion

AFT records is compared with OpenAlex author demographics, which reflect the widest possible researcher population who ever published any global report. Fig4 a. shows over- or under- representation of the countries author, namely the relative registration ratio of the county compared to the share of researchers in the world. The country of the author was inferred from the location of author's registered institution. US, UK, and French colonial institutions have higher registration rate than other countries, among other well represented developed countries in Europe. Similar disparity is between disciplines (fig4 b). Although the disproportionately high neuroscience representation is due to that the ser vice started in the discipline and accumulated most effort. Researchers from psychology, biochemistry have a higher registration rate than average, followed by nursing, medicine and immunology, which may reflect the disciplinal prox imity. Furthermore, the registered researchers are renowned researchers; they have higher mean impact and productivity, with median 2.3×10^2 and 3.2×10^4 times larger than the average, respectively(fig.4 d). Note that the both y axis for impact and productivity are log scaled. Surprisingly, nonetheless the registered researcher does not have significantly different academic age, which is consistent through all the cohort of the career start year (fig.4 c). Gender imbalance is slightly higher than global average

(fig.4 e). Expected ratio is calculated from the weighted mean of inferred degree of the registered researchers.

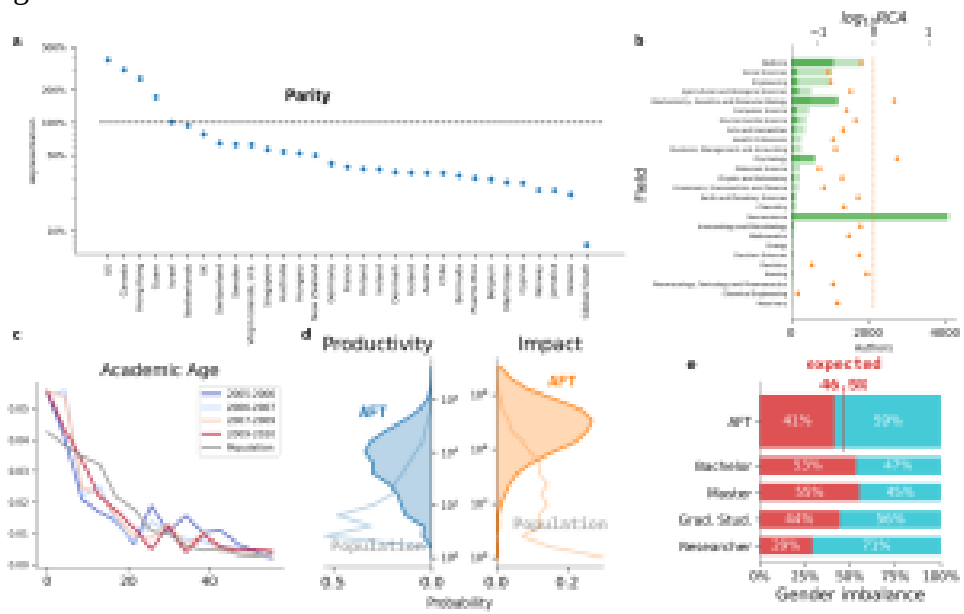

Figure 4. demographics of AFT. a) Top 30 most represented countries in AFT. b) Representation difference by fields. c) Academic age demographics, strati fied by the registration year cohort. d) Distribution of authors with a certain productivity and impact. e) Registered authors gender balance. Following bars shows the global imbalance by degree.

Academic Family Tree is a community-supported registration server that covers researchers and their mentor-trainee relationship from various background. Although it has some degree of registration bias, academic genealogy can yield a rich information on the knowledge flow, if dealt with an adequate calibrations.

Acknowledgements

Stephen David gave us a thorough instruction on academic family tree data format interpretation.

References

Katy B¨orner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans. Skill discrepancies be tween research, education, and jobs reveal the critical need to supply soft skills for the data economy. 115(50):12630–12637.

Nicolas Carayol and Thuc Thi. Why do academic scientists engage in inter disciplinary research? 14(1):70–79. Publisher: Oxford University Press.

Stephen V. David and Benjamin Y. Hayden. Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. 7(10):e46608. Publisher: Public Library of Science.

James A. Evans. Industry collaboration, scientific sharing, and the dissem ination of knowledge. 40(5):757–791.

National Institutes of Health and others. Biomedical research workforce working group report.
Qing Ke, Lizhen Liang, Ying Ding, Stephen V. David, and Daniel E. Acuna. A dataset of mentorship in bioscience with semantic and demographic esti mations. 9(1):467. Publisher: Nature Publishing Group.
Katia Levecque, Frederik Anseel, Alain De Beuckelaer, Johan Van der Hey den, and Lydia Gisle. Work organization and mental health problems in PhD students. 46(4):868–879.
Cassidy R. Sugimoto, Chaoqun Ni, Terrell G. Russell, and Brenna Bychowski. Academic genealogy as an indicator of in terdisciplinarity: An examination of dissertation networks in library and information science. 62(9):1808–1828. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21568.

Supplementary

Data Availability

Our data and code will available at our project repository. The matched ID and other datasets will be uploaded on Zenodo as well.

Data Normalization

Name normalization

Some of the records have non-English names, nicknames, and other supplementary names, all of which we could observe were parenthe sized. We store those records separately in"rawname oaid.csv". Punctuation marks and other special latin characters are not modified. One record on AFT has ORCID while both firsname and lastname are missing (pid=944562). We took this and matched to openalex. On the other hand, two records on AFT has middlename while both firsname and lastname are missing (pid=879367, 929462). As these records as unreliable, we ignored this record throughout the process.

ORCID

Some of the ORCIDs are recorded on AFT while it is not disclosed on ORCID as a public record, result in no match on OpenAlex.

Semantic Scholar ID (s2id)

AFT dataset does have an id column to store semantic scholar ids, which is the numerical string at the end of the URL in the semantic scholar profile page and can be retrieved via API. Semantic Scholar has 79 million author records (viewed on Jan. 18, 2025) which is comparable to openAlex (101 million). We did not use them to match authors because 1. ORCID is a nonproprietary while S2ID is not, 2. OpenAlex does not currently support semantic scholar id in their database.