REPRODUCTION OF SCIENTISTS

Seokkyun Woo [a,*], and Sotaro Shibayama [b]

[a] KAIST, Graduate Schol of Science and Technology Policy

[b] The University of Tokyo, Institute for Future Initiative, Tokyo, Japan

[*] Corresponding author: wsk618@kaist.ac.kr

Abstract

Academic family lineages, consisting of a chain of mentor-mentee relationships, offers the critical path through which young scientists are trained and scientific knowledge is passed down. Academic lineages are subject to competition – failing to win the competition can terminate the lineage, shrinking the line of research and giving space away to others. Such evolutionary fates are determined by scientists' various strategies in navigating the scientific space. Among others, this study focuses on the *exploration* of mentees' topics during the academic training period as a critical strategic decision that determines the evolution of academic lineages. Drawing on a novel dataset linking 17,011 U.S. PhD mentors and 78,102 mentees to large-scale bibliometric records, this study provides two sets of findings. First, the result highlights the incompatibility of various family-level goals – for one, exploration is detrimental to the mentor's productivity but is desirable to expand the family's semantic territory. This conflict becomes pronounced in large families as well as in fast-growing research fields, whereas refraining from exploration becomes the optimal strategy for all considered goals in small families and in established fields. Second, our mentee-level analyses help understand the underlying mechanisms – namely, exploration during PhD training is on average inefficient for individual mentees' productivity but makes a lasting impact on their topic selection, leading to a greater semantic coverage of the family; mentees in a large family compete one another for employment opportunities, resulting in earlier exit, which is mitigated by exploration; and this advantage of exploration is particularly noticeable in fast-growing fields.

Keywords: Academic Genealogy, Doctoral Training, Sociology of Science

1. Introduction

As science underpins our modern knowledge-based society, it is essential that its foundational resource—scientists—be sustainably reproduced (Bozeman & Corley, 2004; Laudel & Gläser, 2008; Stephan, 2012). In this regard, an academic family lineage, consisting of a chain of mentor-mentee relationships, offers the critical path through which young scientists are trained, scientific knowledge is passed down, and thereby scientific communities are sustained (David & Hayden, 2012; Dores, Benevenuto, Laender, & Ieee, 2016; Laudel & Gläser, 2008; Malmgren, Ottino, & Amaral, 2010; Shibayama, 2019).

Importantly, academic lineages are subject to competition – failing to win the competition can potentially terminate the lineage, shrinking the line of research and giving space away to other lineages. Such evolutionary fates are determined by scientists' strategies in navigating and exploring the scientific space. In particular, in the nascent stage of academic life, young mentee scientists (typically PhD students) are engaged in certain research topics under the guidance of their mentors, which influences the trajectories of their research in the long term. Here, the topics may be selected around the mentor's neighborhood or may be explored far away from the mentor's expertise. These options have pros and cons – the former may be safe but can lead to overcrowding, whereas the latter may lead to untapped opportunities but can be inefficient. An analogy can be drawn from population dynamics, where 'dispersal'—the geographical distance between the sites of parents and offspring—is a key strategic factor (Hamilton & May, 1977), where the optimal level of dispersal is subject to various environmental conditions. Similarly, we argue that exploration of mentees' topics is a critical strategic decision that determines the evolution of academic lineages. Here, a stereotypical assumption is that they usually follow their mentors' research agendas with limited exploration (Delamont et al., 1997), but closer examination reveals variation in the semantic distance between mentors' and mentees' research topics during PhD training (Shibayama, Baba, & Walsh, 2015). In this study, we inquire how this strategy affects an academic family's long-term outcomes such as productivity, survival, and semantic territory, and how it varies across various environments.

This is an inquiry situated at the cusp of a few literatures. On the one hand, the academic genealogy literature has helped us understand the evolution of academic families by tracing the lineage among scientists typically through mentor-mentee relations in doctoral education (David & Hayden, 2012; Dores et al., 2016; Malmgren et al., 2010; Rossi, Freire, & Mena-Chalco, 2017). While the literature illustrates the topological features of lineage evolution and their temporal dynamics (Damaceno, Rossi, Mugnaini, & Mena-Chalco, 2019; David & Hayden, 2012; Malmgren et al., 2010; Rossi, Damaceno, Freire, Bechara, & Mena-Chalco, 2018), it remains largely descriptive and disconnected from the micro-foundations underlying the evolution. On the other hand, the higher education and science studies literatures investigate the practices of research training, including the exploration of mentees' topics (Shibayama, 2019; Wang & Shibayama, 2022). While the literature has revealed substantial heterogeneity in the practices in the postgraduate training context (Bastalich, 2017; BROWN & ATKINS, 1988; Hockey, 1991; Marsh, Rowe, & Martin, 2002), it has paid limited attention to the

consequences of those practices, especially at the academic family level. This study seeks to bridge these gaps.

To this end, we investigate how topic exploration determines long-term outcomes of academic lineages, drawing on a newly constructed dataset that links U.S. doctoral dissertation records to large-scale bibliometric data of 17,011 mentors and 78,102 mentees. Our findings are twofold. First, the empirical analyses highlight the incompatibility of various family-level goals – for one, exploration is detrimental to the mentor's productivity but is desirable to expand the family's semantic territory. This conflict becomes pronounced in large families as well as in fast-growing research fields, whereas refraining from exploration is the optimal strategy for all considered goals in small families and in established fields. Second, our mentee-level analyses help understand the underlying mechanisms – namely, exploration during PhD training is on average inefficient for individual mentees' productivity but makes a lasting impact on their topic selection, leading to a greater semantic coverage of the family; mentees in a large family compete one another for employment opportunities, resulting in earlier exit, which is mitigated by exploration; and this advantage of exploration is particularly noticeable in fast-growing fields.

2. Data

2.1 ProQuest Dissertations and Theses Global (PQDT Global) data

We use ProQuest Dissertations & Theses Global (PQDT Global) to construct our main empirical dataset, with the identification of mentor and mentee inferred from doctoral dissertation authors and their supervisors. PQDT Global provides structured metadata of millions of dissertations, including record titles, abstracts, year of completion, degree type, granting institution, fields of study, and supervisor names when available. To ensure accurate genealogical reconstruction, we limit our primary data by: 1) removing non-PhD degrees (such as master's degrees or doctorates of education), 2) excluding dissertation records from non-US institutions, and 3) excluding records where supervisors are not listed as previous dissertation authors in PQDT Global. The final criterion specifically enables us to trace academic lineages starting from the supervisor's own PhD and prevents left-censoring in our analysis. To assess the quality of our dataset, particularly in terms of how well it captures the actual number of PhDs produced by US academic institutions, we compared the number of PhDs recorded in our dataset with figures from the NSF's Survey of Earned Doctorates (SED). As shown in the Fig. A1.C in the Appendix, our dissertation data closely follow the SED data, with only minor discrepancies in a few years where our dataset has some missing records.

2.2 Bibliometric Data

To identify knowledge production activities, we linked publication records to all authors in our mentor-mentee pair dataset. Since all supervisors (mentors) in our data must appear as student authors in our dissertation records, our objective was to match each dissertation author, both students and supervisors, to their respective publication records. For publication data, we used

SciSciNet(Lin, Yin, Liu, & Wang, 2023), a large-scale open-access bibliometric dataset built on the former Microsoft Academic Graph (MAG). SciSciNet provides detailed metadata for over 100 million publications, including disambiguated author identifiers, institutional affiliations, publication year, titles, and fields.

2.3 Construction of Mentor–Mentee Links

PQDT Global provides supervisor names for the majority of US PhD dissertations, with coverage ranging from 80 to 90% in the late 1980s to late 1990s, and exceeding 95% from the late 1990s onward (Fig.A1.D in the Appendix). Among dissertation records that report supervisor information, approximately 90% list a single supervisor, while the remaining 10% list two or more supervisors (Fig.A2.A in the Appendix). For dissertations with multiple supervisors, we use the first-listed supervisor as the primary supervisor. Comparison of the observed frequency of alphabetical ordering with the expectation under random ordering indicates a modest excess of alphabetical ordering in PQDT Global, suggesting some degree of intentional ordering, though the magnitude of the deviation from chance is small (Fig.A2.B in the Appendix). While mentor–mentee relationships can be directly inferred by linking student names and supervisor names within dissertation records, a key challenge is identifying student authors who later appear as supervisors in subsequent dissertations. Resolving this challenge allows us to construct academic family lineages and to measure the number of students trained by former mentees once they become supervisors. We address this by identifying individuals who appear both as student authors and, in later years, as supervisor authors in PQDT Global. We implement a multi-step matching strategy to link student and supervisor records belonging to the same individual.

2.3.1 Disambiguating Student Authors and Supervisors.

The first necessary condition for matching is that a student author and a supervisor author share the same name (Fig.A3.A in the Appendix). We therefore standardized all names into first–last name blocks and identified candidate pairs with identical name blocks. We then applied additional rules based on middle-name information. First, we removed pairs with conflicting middle names. For the remaining pairs, we assigned similarity scores based on the degree of middle-name agreement. Exact matches in middle names received the highest score (3 points). Partial matches—where one record lists a full middle name and the other lists only a matching initial—received 2 points. Cases in which one record includes a middle name while the other does not received 1 point. We further incorporated name rarity, based on the idea that rare name pairs are more likely to correspond to the same individual. We assigned 1 additional point to name blocks that appear five times or fewer in the dataset and subtracted 1 point for very common name blocks, defined as those in the top 5% of the frequency distribution (50 or more occurrences).

The second criterion ensures the temporal plausibility of the mentor–mentee relationship. We assume that a student cannot complete a dissertation before their supervisor's own dissertation year and that a minimum of five years must elapse between a supervisor's PhD completion and a student's dissertation. This threshold reflects the average duration of doctoral training, which

is around 5 to 6 years. We therefore excluded candidate pairs in which the student's dissertation year falls within five years of the supervisor's dissertation year. We further assigned 2 points to pairs with an interval between 8 and 25 years, reflecting a typical academic career trajectory (need to cite and be based on the evidence), and 1 point to intervals between 5 and 8 years. Implausibly short or long intervals (less than 5 years or more than 25 years) were penalized by 2 points. In addition to name matching and PhD timing, we incorporated information on dissertation fields of study. We measured field similarity between students and potential supervisors by embedding dissertation field descriptions into a dense vector space using a BERT-based language model (Singh, D'Arcy, Cohan, Downey, & Feldman, 2022) and computing cosine similarity between the resulting embeddings. We assigned 2 points for high similarity (≥ 0.95), 1 point for moderate similarity (≥ 0.90), and penalized low similarity (< 0.90) with –2 points. We summarize the matching strategy and associated scoring rules in Table A1 in the Appendix.

2.3.2 Final Assignment

After computing match scores for all candidate mentor–mentee pairs, we sorted candidates by student and retained only the highest-scoring supervisor for each student, resulting in a one-to-one mentor–mentee linkage. Under this approach, each mentee is assigned to a single mentor, while each mentor may be matched to multiple mentees. Using this procedure, we identified 54,980 mentors who had previously appeared as students in the dissertation records and matched them to 196,495 unique mentees, yielding 196,495 unique mentee–mentor pairs. In our analysis, we restricted our sample to 48,746 mentors and their respective 168,647 mentee-mentor pairs, focusing on pairs with high match scores (Fig.A4.A in the Appendix). In our matched mentor-mentee dataset, mentors, on average, have appeared as a supervisor on 3.46 mentees' dissertations, with 2 being the median value. In other words, most mentors produce 2 to 3 students, while there are few highly productive mentors, around the top 5% in the distribution, who have produced more than ten mentees (Fig.A4.B in the Appendix).

2.4 Linking PQDT Global and SciSciNet

We linked publication records to individuals identified in our mentor–mentee pairs. Because all mentors in these pairs also appear as student authors in PQDT Global, our objective was to match each dissertation author—both mentees and mentors—to their corresponding publication records in SciSciNet (Fig.A3B in the Appendix). This allows us to connect dissertation-based academic genealogies to downstream publication activity.

2.4.1 Selecting Candidate Pairs

We first standardized author names in both PQDT Global and SciSciNet into first–last name blocks and retained only candidate pairs with identical name blocks. To ensure temporal plausibility and reduce false matches, we excluded SciSciNet publications whose publication years fell outside a ±5-year window around the dissertation year. This window captures both early publications around degree completion and short publication lags while substantially reducing the candidate search space.

2.4.2 Scoring and Matching

For each candidate dissertation–publication author pair, we computed a composite match score by summing points across multiple criteria (Table A2 in the Appendix), similar to what we did for matching mentor and mentee within PTDT Global data. Firstly, middle-name similarity was scored as follows: exact middle-name matches received +3 points, partial matches—where one record lists a full middle name and the other lists only a matching initial—received +2 points, cases in which one record lists a middle name and the other does not received +1 point, and conflicting middle names were penalized by –2 points. Name rarity was incorporated based on the frequency of first–last name blocks in the dataset: name blocks appearing five times or fewer received +2 points, while very common name blocks (50 or more occurrences) were penalized by –1 point. We also consider the time gap between dissertations and publications. Candidate pairs receive +2 points if the matched publication falls within ±5 years of the dissertation year. Pairs that did not satisfy this condition received no additional points. Next, we evaluated semantic similarity between PQDT Global records and SciSciNet records. Title similarity was measured using cosine similarity of SPECTER2 embeddings (Singh et al., 2022), with similarities ≥0.98 receiving +3 points, similarities ≥0.90 receiving +2 points, similarities <0.72 (median value) receiving –1 point, and intermediate values receiving 0 points.

Field similarity was measured by mapping both PQDT Global and SciSciNet field classifications into the SPECTER2 embedding space. We then computed cosine similarities and assigned higher scores to pairs that share similar fields in the embedding space. Specifically, we assigned +2 points for similarities ≥0.98, +1 point for similarities ≥0.95, and –1 point for similarities <0.95. Institutional similarity was evaluated where affiliation information was available: cosine similarity ≥0.98 received +2 points. We gave an additional +2 points if PQDT authors and SciSciNet authors shared the same institution (SPECTER2 cosine similarity >0.98). Lastly, we assigned an additional +1 point when both the PQDT Global record and the SciSciNet author were affiliated with a university. The total match score is the sum of points across all criteria, with higher scores indicating stronger evidence that a PQDT Global author and a SciSciNet publication author correspond to the same individual.

2.4.3 Final Assignment

To finalize the matching, we selected, for each PQDT Global author, the top-scoring SciSciNet author. This produced 741,302 unique PQDT Global authors matched to SciSciNet authors. Because a few cases involve more than one PQDT Global author matched to a single SciSciNet author, we removed duplicates by keeping only the highest-scoring PQDT Global author for each shared SciSciNet author. This left us with 720,657 unique PQDT Global authors. For comparison, we began with 1,059,244 US PhD dissertation records from PQDT Global dataset. This suggests that, at a minimum, approximately 68% of PhD graduates have their doctoral dissertations published (Fig.A5.A in the Appendix).

2.5 Combining mentor–mentee pairs from PQDT Global with SciSciNet author matches

We combine two datasets: mentor–mentee pairs identified from the PQDT Global dissertation records and matched PQDT Global author–SciSciNet author pairs. Merging these datasets

allows us to observe both mentor–mentee relationships and the subsequent publication trajectories of individuals. The PQDT Global data contain 168,647 mentor–mentee pairs, comprising 168,647 mentees and 48,746 mentors. Because some mentees later appear as mentors themselves, this structure yields 212,636 unique PQDT Global authors. We then link these authors to SciSciNet using the PQDT Global author–SciSciNet author matching dataset. Of the 212,636 unique PQDT Global authors, 170,957 are successfully matched to SciSciNet records, corresponding to a coverage rate of 80.4%. This linkage enables us to track the publication histories of the majority of mentors and mentees in our sample using comprehensive bibliometric data.

3. Methods

3.1. Measuring Exploration

3.1.1 Mentee's Semantic Distance

We operationalize mentors' exploration strategies by aggregating computed topical distance between a focal mentor's body of work and his or her mentees' dissertations. Specifically, we embed the titles and abstracts of mentees' dissertations and the titles of mentors' publications into a shared semantic space and compute pairwise distances between them. The embeddings are generated using a Sentence-BERT model (Reimers & Gurevych, 2019). To ensure the temporal relevance of a focal mentee's dissertation, we restrict the mentors' publications to those published within a 11-year window preceding and including the mentee's dissertation year. We also removed the mentor's publications co-authored with mentees in order to avoid mechanical correlation between the exploration measure and the mentee's knowledge production activities.

We define the mentee's semantic distance as the median of 1 minus the cosine similarity between the mentee's dissertation and the selected mentors' publications:

- $d_i = median_{j=1}^{N_{g(i)}} \left(1 - cos\left(Dissertation_i, Paper_{g(i),j}\right)\right)$

where d_i is the median topical distance for mentee i, $\mathrm{Dissertation_i}$ is the embedding of mentee i's dissertation, and $\mathrm{Paper}_{g(i),j}$ denotes the embedding of j-th paper published by mentee i's mentor *g(i)* within the 11-year window. $N_{g(i)}$ is the number of such publications.

3.1.2 Mentor's Exploration Strategy

After computing the semantic distance between each mentor and their respective mentee, we aggregate this measure at the mentor level to operationalize the mentor's exploration strategy. For each mentor g, we calculate the average of the mentee's median distances:

- $\mathrm{Exploration}_g = \frac{1}{n_g} \sum_{i \in S_g} d_i$

where Exploration$_g$ captures the exploration of the mentor g, S_g is the set of mentees supervised by mentor g. Lastly, the mentor family size n_g is the number of mentees trained by a given mentor, which serves as a proxy for the scale of the mentor's academic group.

Of the 168,647 mentor-mentee pairs (corresponding to 168,647 unique mentees), we computed a semantic distance measure for 134,638 mentees (Fig. A6.B in the Appendix). The reduction in sample size is driven by two exclusions: we omitted mentors' publications co-authored with the focal mentees and dropped mentees whose mentors had no publications within the selected 11-year time window. Using these 134,638 mentee-level semantic distances, we constructed a mentor-level exploration measure by averaging across all mentees associated with each mentor. This aggregation yielded 39,117 unique mentors with exploration measures (Fig.A6.B in the Appendix).

3.2. Field Dynamics

3.2.1 Field-level topic turnover

We measure field dynamics using a topic turnover metric constructed at the subfield-year level. The rationale is that a field's intellectual content can be approximated using discrete linguistic units. Specifically, by extracting noun phrases from publication titles, we can track the introduction of new topics and the disappearance of older ones over time (Cheng et al., 2023; Milojević, 2015). The extent to which new noun phrases enter and previously used ones are replaced provides a direct indicator of how rapidly a field's intellectual focus evolves. For each of 293 subfields in SciSciNet and each year between 1991 and 2022, we extract sets of noun-phrases topics from titles of publications assigned to that subfield.

For a given subfield *j*, and year *t*, we define:

- $B_{j,t}$: the set of noun-phrases appearing in year *t*–1.
- $A_{j,t}^{cum}$: the cumulative set of topics appearing in the same subfield from 1950 through *t*–2.

We then computed cumulative topic turnover as:

- $Turnover_{j,t}^{cum} = 1 - \frac{\left|A_{j,t}^{cum} \cap B_{j,t}\right|}{\left|A_{j,t}^{cum} \cup B_{j,t}\right|}$

This measure captures the extent to which recently used topics (as proxied by noun phrases) differ from the historical topic of the field. Higher values indicate faster topic replacement. Although topic turnover is computed annually at the subfield level, our analyses use a time-invariant measure of field dynamics. Our decision is motivated by the empirical pattern that cumulative topic turnover exhibits relatively little variation over time within a subfield, while differing substantially across subfields. Thus, we obtain a time-invariant subfield-level measure by averaging cumulative turnover across all available years for each subfield.

3.2.2 Mapping field dynamics to papers and mentors

Because individual papers may be assigned to multiple subfields, we map field-level dynamics to the paper level using a weighted average across subfields, with equal weights assigned to each associated subfield. Once all papers were assigned field dynamic measures, we aggregated them at the mentor level by taking the average of the field dynamics across all papers authored by the mentor. This procedure provides a mentor-level measure of exposure to field dynamism, capturing the average rate of cumulative topic turnover in the knowledge production environments in which the mentor conducts research and trains mentees.

3.3 Measurement of Semantic Territory

3.3.1 Mentor's Semantic Territory

We operationalize the expansion of semantic territory with two complementary measures. The first attempts to capture the full breadth of topics explored within an academic family by leveraging text information from the titles of all papers authored by the mentor and their mentees. We embedded these titles into a shared vector space using a Sentence-BERT model (Reimers & Gurevych, 2019). Then, for each academic family, we computed pairwise cosine similarities among all embedded papers and defined the family-level semantic territory as the minimum observed similarity across pairs of papers (Fig.A7.A in the Appendix). Larger distance values indicate broader semantic territory.

Alternatively, we measure semantic territory using the total number of distinct topics covered by an academic family. We rely on topic labels assigned to each paper by OpenAlex, which are generated by a pre-trained machine learning model developed by OpenAlex and CWTS that classifies publications based on citation patterns and textual features[1]. Using these topic assignments, we aggregate all papers authored by the mentor and their mentees and count the number of unique topics represented within each academic family. Larger values indicate broader semantic territory (see Fig.A7.B in the Appendix).

3.3.1 Mentee's Semantic Territory

We operationalized the expansion of semantic territory at the mentee-level using the same two complementary measures as in the mentor-level analysis. Instead of using publications from the entire academic family, the mentee-level measures are constructed using only the mentee's own publications. Specifically, we measure semantic territory based on (i) embedding-based distances (Fig.A7.C in the Appendix) and (ii) OpenAlex topic counts.

[1] https://www.leidenmadtrics.nl/articles/an-open-approach-for-classifying-research-publications; https://help.openalex.org/hc/en-us/articles/24736129405719-Topics

3.4 Econometric Specifications

3.4.1. Mentor-level Specification

We estimate the variations of the following regression specification to investigate how exploration influences the three key dimensions of the family-level goals and how these relationships are moderated by family size and field dynamics.

$$E[Y_j|X_j] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + Z_j^{'}\delta$$

$$E[Y_j|X_j] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + \beta_4 \text{Exploration}_j \times \text{FamilySize}_j + Z_j^{'}\delta$$

$$E[Y_j|X_j] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + \beta_4 \text{Exploration}_j \times \text{FieldDynamics}_j + Z_j^{'}\delta$$

Y_{j} denotes our primary outcome of interest for mentor j, which includes (1) the mentor's own productivity (total number of publications at the career-level), (2) the sustainability of the family lineage (probability of producing future mentors), and (3) the expansion of the semantic territory (coverage of semantic space by the family). Our key explanatory variable, exploration, $(exploration_j)$ measures the extent to which a mentor allows their students to produce dissertations that deviate from their own primary research areas. To examine the differential effects of exploration, we moderate the exploration variable with family size ($FamilySize_j$) and field dynamics ($FieldDynamics_j$) of mentors' primary research fields. To account for unobserved heterogeneities across mentors, we include a vector of control variables, Z_i', including mentors' average citation impact and whether they have graduated from elite universities (Ivy Plus Universities[2]). They also include cohort fixed effects based on the mentor's PhD dissertation year, accounting for potential temporal differences across generations, as well as field fixed effects to control for discipline-specific baseline differences. Fields were derived from the most common OpenAlex field categories in the mentor's publications.

3.4.2 Mentee-level Specification

For the mentee-level analysis, we estimate mentor-level random effect models to account for the nested structure of the data, in which mentees are clustered within mentors. We denote Y_{ij} as the outcome for mentee i, trained by mentor j. We examine three primary outcomes: (1) mentee productivity (total number of publications at the career-level); (2) career survival (publishing career longevity); and (3) expansion of the semantic territory (coverage of semantic space by the mentee).

$$E[Y_{ij}|X_{ij}] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + Z_{ij}^{'}\delta$$

$$E[Y_{ij}|X_{ij}] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + \beta_4 \text{Exploration}_j \times \text{FamilySize}_j + Z_{ij}^{'}\delta$$

$$E[Y_{ij}|X_{ij}] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + \beta_4 \text{Exploration}_j \times \text{FieldDynamics}_j + Z_{ij}^{'}\delta$$

[2] https://en.wikipedia.org/wiki/Ivy_Plus

The key explanatory variables are mentor-level measures of exploration, family size, and field dynamics. As control variables, we include indicators for whether the mentor obtained a PhD from an elite university and whether the mentee obtained a PhD from an elite university, the mentor's average citation impact, the mentee's PhD graduation year, and the mentee's primary field area.

4. Results

First, we examine the degree to which mentors allow their mentees to engage in research topics remote from their own. Fig.1 shows the distribution of mentor-mentee topic distances during PhD training. As expected, mentor-mentee pairs tend to engage in similar topics[3] but some mentees engage in remote topics. We inquire how such different levels of exploration affect multiple family-level outcomes that mentors may be interested in pursuing. For one, scientists in a mentorship role tend to be under fierce competition, which can make them prioritize their own productivity, potentially at the cost of their mentees' (Shibayama, 2019). Nonetheless, mentors may consider doctoral education as a moral obligation to the academic community (Hockey, 1996), placing mentees' interests above their own. Some mentors may view the preservation of their research lines as a primary goal, while others aspire for their research to branch out and expand into diverse directions. Thus, we study as family-level outcomes (1) the mentor's own productivity (irrespective of the fate of the family), (2) the survival of the family lineage, and (3) the expansion of the semantic territory of the family. Note that these goals are inter-related and can be in conflict.

In what follows, we first test the effect of exploration on the family-level outcomes, which we further probe from two angles. First, we analyze how the effect of exploration changes across different contexts. In particular, we highlight two contextual factors – 1) family size, or the number of mentees supervised by a mentor (Fig.2), and 2) topic turnover rate, or the rate at which research topics in a field are replaced (Fig.3) – which are likely to change the rationale of exploratory strategy. Second, on top of family-level outcomes, we also make analyses at the individual mentee's level to elucidate the underlying mechanisms.

4.1 Exploration and family outcome.

Fig.4 visualizes the predicted relationships between exploration and the mentor's publication productivity (Fig.4A–C), the number of future mentors produced (Fig.4D–F), and the semantic territory covered (Fig.4G–I). Estimates are based on regression specifications reported in Table 1–3. On average, greater allowance for exploration is negatively associated with mentors' publication productivity (Fig.4A), indicating that mentors who permit mentees to pursue topics that deviate from their own primary research lines tend to publish less than mentors who direct mentees to remain closely aligned with their own research. Meanwhile, this negative relationship did not differ significantly across family size (Fig.4B) and field dynamics (Fig.4C).

[3] *Exploration* is defined as the semantic distance between mentees' dissertation and mentors' publications and is measured by the cosine distance (1 – cosine similarity) between the document embeddings. The mean of exploration is 0.711, suggesting reasonable similarity between the pairs.

Regarding the continuation of the academic lineage, we find no association between the exploration strategy and the number of future mentors, on average (Fig.4D). This null relationship holds for both large and small families (Fig.4E). Meanwhile, we find that field dynamics condition this relationship. Mentors operating in relatively fast-evolving fields experience a positive association between exploration and the production of future mentors, despite they are, on average, less likely to produce future mentors, on average, than those operating in slow-evolving fields (Fig.4F). Thus, in rapidly evolving fields, allowing mentees explore topics distant from their own is associated with a higher likelihood of reproducing the mentor's academic lineage.

When it comes to the semantic territory covered by the academic family, we find that exploration is, on average, positively associated with greater territory coverage (Fig.4G). In other words, mentors who allow mentees to pursue topics that deviate from their core research areas tend to cover a broader semantic territory than those who do not. This relationship, however, differs by family size. Among mentors with large families, exploration is positively associated with semantic territory coverage, whereas among mentors with small families, the association is negative (Fig.4H). Field dynamics also condition this relationship. Mentors operating in fast-evolving fields exhibit a stronger positive association between exploration and the semantic territory covered (Fig.4I).

The result highlights trade-offs between these goals in determining the degree of exploration. On average, exploration appears detrimental to mentors' productivity but preferable for families' semantic territory. The contextual analyses suggest that this trade-off becomes pronounced in large families as well as in fast-growing research fields. Specifically, while exploration compromises mentors' productivity, it increases families' semantic coverage when the cohort size is larger than 4 or when the field dynamics is around its 55^{th} percentile value. On the other hand, in smaller families and in established fields, all three goals are compatible by refraining from exploration.

4.2 Exploration and Mentee Outcomes.

To further elucidate the underlying mechanisms, we examine the relationship between exploration and mentee-level outcomes – mentees' publication productivity, career dropout rates, and semantic coverage, respectively corresponding to the family-level outcomes.

Regarding academic lineage, we find that exploration is, on average, negatively associated with career longevity, as indicated by higher career exit hazards with greater exploration (Fig.5D). While mentees from larger families generally face higher career exit hazards than those from smaller families, the association between exploration and the hazard rate is substantially stronger among mentees from smaller families than among those from larger families (Fig.5E). Meanwhile, mentees in fast-evolving fields experience a negative association between exploration and the career exit hazard, whereas those in slow-evolving fields experience a positive association (Fig.5F). In other words, exploration may be a favorable strategy for

mentees trained in fast-evolving fields, while it is the opposite for mentees from slow-evolving fields.

Next, we analyze the relationship between mentor's exploration strategy and individual mentees' publication productivity after controlling for career longevity. We then find that exploration is negatively associated with mentees' publication productivity (Fig.5A). Further when conditioned on family size, exploration is negatively associated with productivity among mentees in smaller families, whereas no association is observed in larger families (Fig.5B), despite the fact that, on average, smaller families tend to produce more productive mentees than larger families. The moderating effect of field dynamics is found to be insignificant (Fig.5C).

Lastly, we examine the effect of exploration on individual mentees' semantic territory with their publication productivity controlled for. The result suggests that, on average, the exploration strategy is positively associated with the semantic territory covered by mentees (Fig.5G). In other words, mentees whose dissertation topics deviate from their mentors' primary research areas tend to cover a broader semantic territory over the course of their careers than those who remain closely aligned with their mentors' core research area. This association varies seem unaffected by family size. For mentees from different family sizes, the relationship between exploration and semantic territory is constant (Fig.5H). Also, field dynamics does not seem to change the association between exploration and semantic territory (Fig.5I).

These results imply a few underlying mechanisms. First, the negative effect of exploration on mentees' productivity (Fig.5A) suggests inefficiency in exploration – i.e., mentees fail to utilize mentors' expertise effectively. This inefficiency appears mitigated in large families and augmented in small families (Fig.5B), perhaps because of greater opportunity of mentee-mentee collaboration. At the family level, the negative effect is observed in mentors' productivity due to limited opportunities of mentor-mentee collaboration (Fig.4A). Second, as to mentees' career longevity, exploration appears associated with earlier drop-out. This negative effect is mitigated in larger families (Fig.5E), which implies that exploration helps reduce job competition between family members. Also, exploration decreasing drop-out in fast-growing fields (Fig.5F) implies that exploration may help mentees find new jobs emerging in such fields. At the family level, this is manifested as a greater number of future mentors from the family in fast-growing fields (Fig.4F). Finally, exploration on average leads to greater semantic coverage at the mentee level (Fig.5G), suggesting that exploration in the early career has a lasting impact on mentees' broader topic selection, and this is translated into the family-level wider semantic territory (Fig.4G). This positive effect is reduced in slow-evolving fields (Fig.4I) and reversed in small families (Fig.4H), possibly because of the aforementioned disadvantage of exploration for career longevity in such conditions (Fig.5EF).

5. Conclusions

Mentor-mentee relationships form academic family lineages, offering a critical path through which scientists and scientific knowledge are reproduced (David & Hayden, 2012; Dores et al., 2016; Laudel & Gläser, 2008; Malmgren et al., 2010; Shibayama, 2019). This study focuses on the exploration of mentees' topics as mentors' key strategy and inquires how it influences the family's long-term outcomes. Our mentor-level analyses highlight the incompatibility among family-level goals – mentor's own productivity as opposed to the family's semantic territory – in deciding the level of exploration. The optimal level depends on the context – exploration is more advantageous in large families and in fast-growing fields. Then, our mentee-level analyses help elucidate underlying mechanisms. Exploration during PhD training is on average inefficient for individual mentees' productivity but makes a lasting impact on the family's semantic coverage. Mentees in a large family compete for employment opportunities, resulting in earlier exit, which is mitigated by exploration, and this advantage of exploration is particularly noticeable in fast-growing fields.

This study contributes to two lines of literature. First, it speaks to academic genealogy literature (David & Hayden, 2012; Dores et al., 2016; Malmgren et al., 2010; Rossi et al., 2017) by demonstrating some micro-foundations behind the evolution of academic lineages. Second, it also contributes to the higher education and science studies literatures (Shibayama, 2019; Wang & Shibayama, 2022) by showing long-term impacts of exploration as a practice in research training.

Future research can extend this work in a few directions. First, while this study focuses on topic exploration as a key strategy, future studies can examine other strategic aspects to understand how different training practices shape family outcomes. Second, richer micro-level data, including lab composition and mentoring styles, would allow to more directly examine the decision-making processes behind exploration and the trade-offs mentors face. Third, empirical analyses in other countries might clarify how different incentives and labor-market structures condition the returns to exploration for both mentors and mentees. Finally, future work could trace intergenerational dynamics beyond the first generation of mentees, examining how exploration-induced semantic breadth propagates—or dissipates—across multiple generations of an academic family, thereby offering a deeper understanding of the long-run evolution of scientific fields.

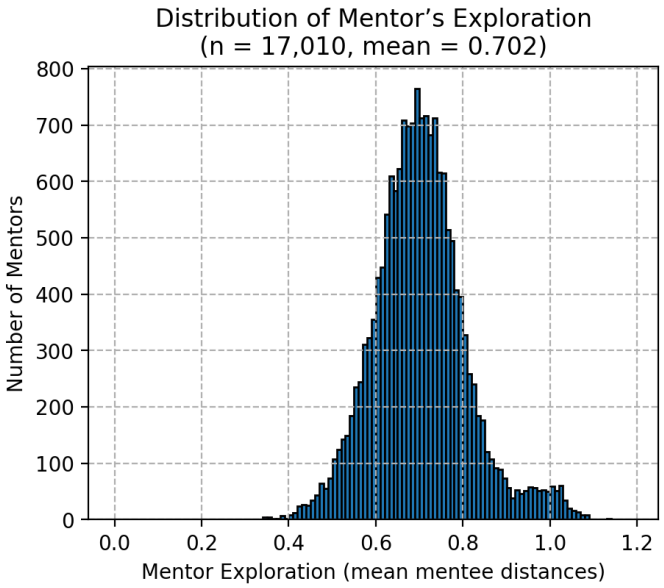

Fig.1 Distribution of Mentor's Exploration. We computed a semantic distance measure between a focal mentor's body of work and their mentees' dissertations. Specifically, we embed the titles and abstracts of mentees' dissertations, as well as the titles of mentors' publications, into a shared semantic space and compute pairwise distances. The embeddings are generated using a Sentence-BERT model. After computing the semantic distance between each mentor and their respective mentee, we aggregate this measure at the mentor level to operationalize the mentor's exploration strategy.

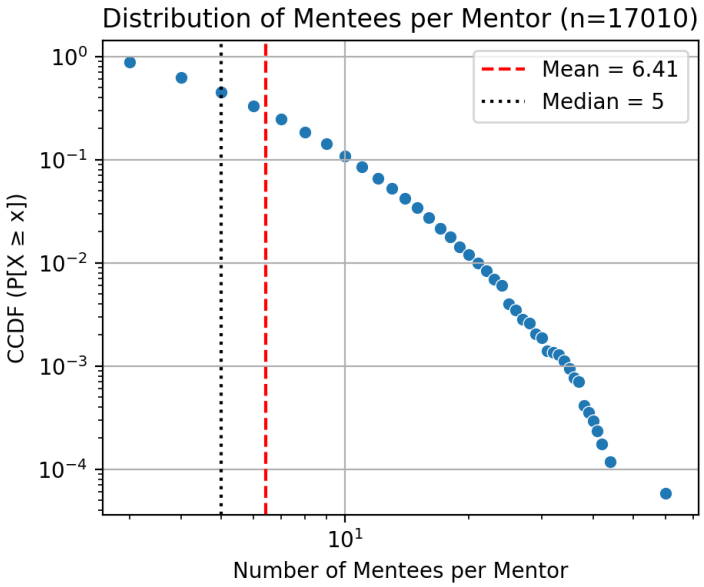

Fig.2 Distribution of Family size. Family size is defined as the number of mentees supervised by each mentor. We identified mentor-mentee links based on PQDT Global data, after careful name disambiguation. Our final sample include 17,010 mentors who have at least three mentees in their publishing careers.

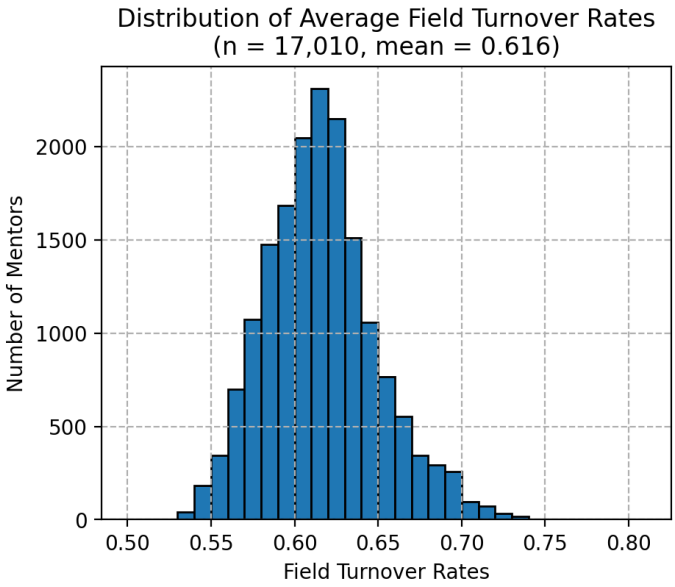

Fig.3 Distribution of Field-level Topic Turnover. We measure the field dynamics using a topic turnover metric constructed at the subfield level. The rationale is that a field's intellectual content can be approximated using discrete linguistic units. By extracting noun phrases from publication titles, we track the introduction of new topics and the disappearance of older ones over time (Cheng et al., 2023; Milojević, 2015). The extent to which new noun phrases enter and previously used ones are replaced provides an indicator of how rapidly a field's intellectual focus evolves. For each of 293 subfields in SciSciNet, we compute the rate at which topics in the subfield are changing over time. Higher values indicate faster topic replacement.

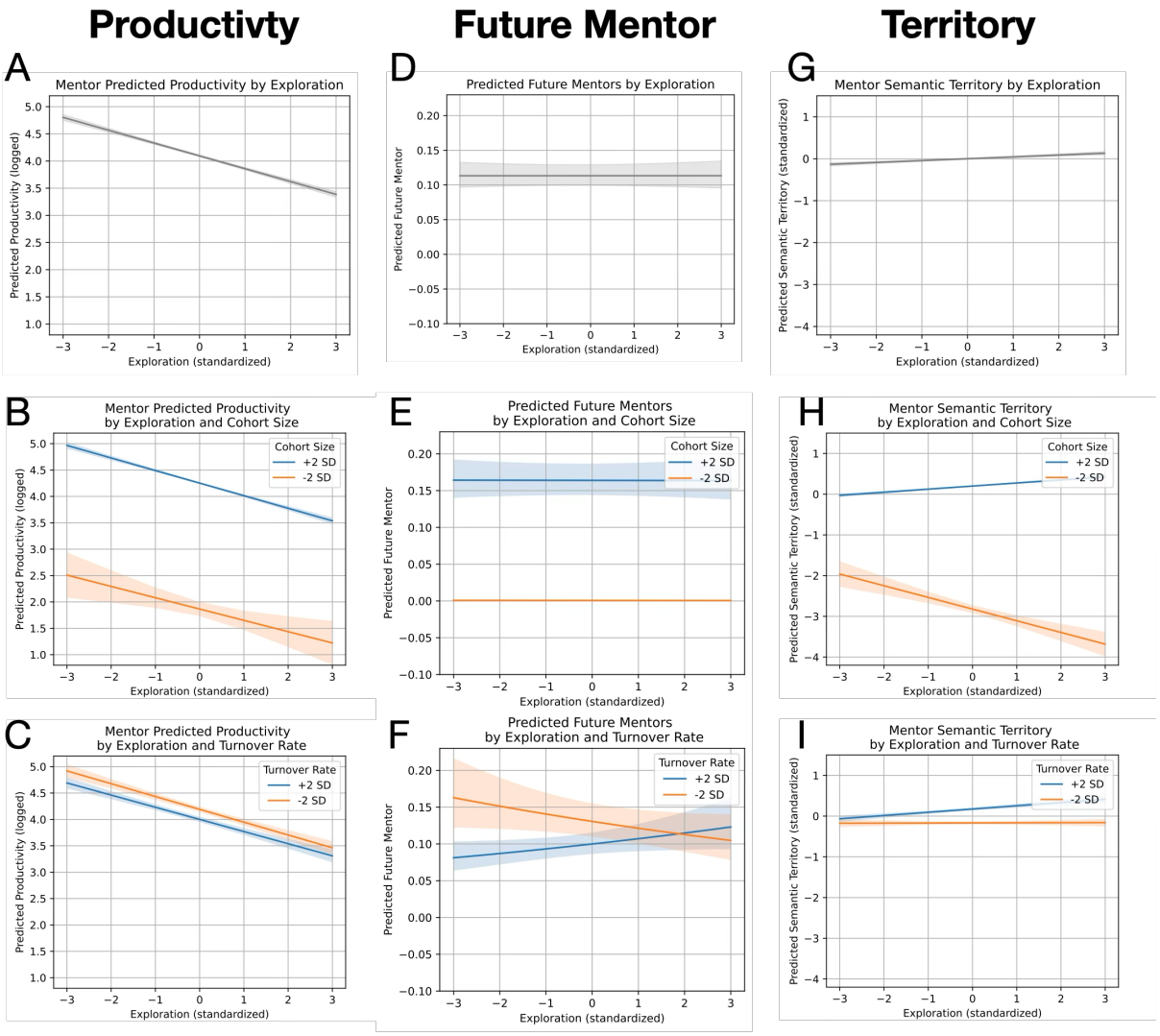

Fig.4 The effects of exploration on mentor-level outcomes. The mentor-level outcomes include total publication productivity (A–C), the number of future mentors produced (D–F), and semantic territory covered (G–I). Effects are shown at average values (top row) and conditional on cohort size (middle row) and turnover rate (bottom row). Blue lines correspond to predicted outcomes at cohort size or turnover rates that are 2 standard deviations above the mean, while orange lines show predictions 2 standard deviations below the mean. Estimates are model-based predictions from Tables 1–3, with shaded area indicating 90% confidence intervals.

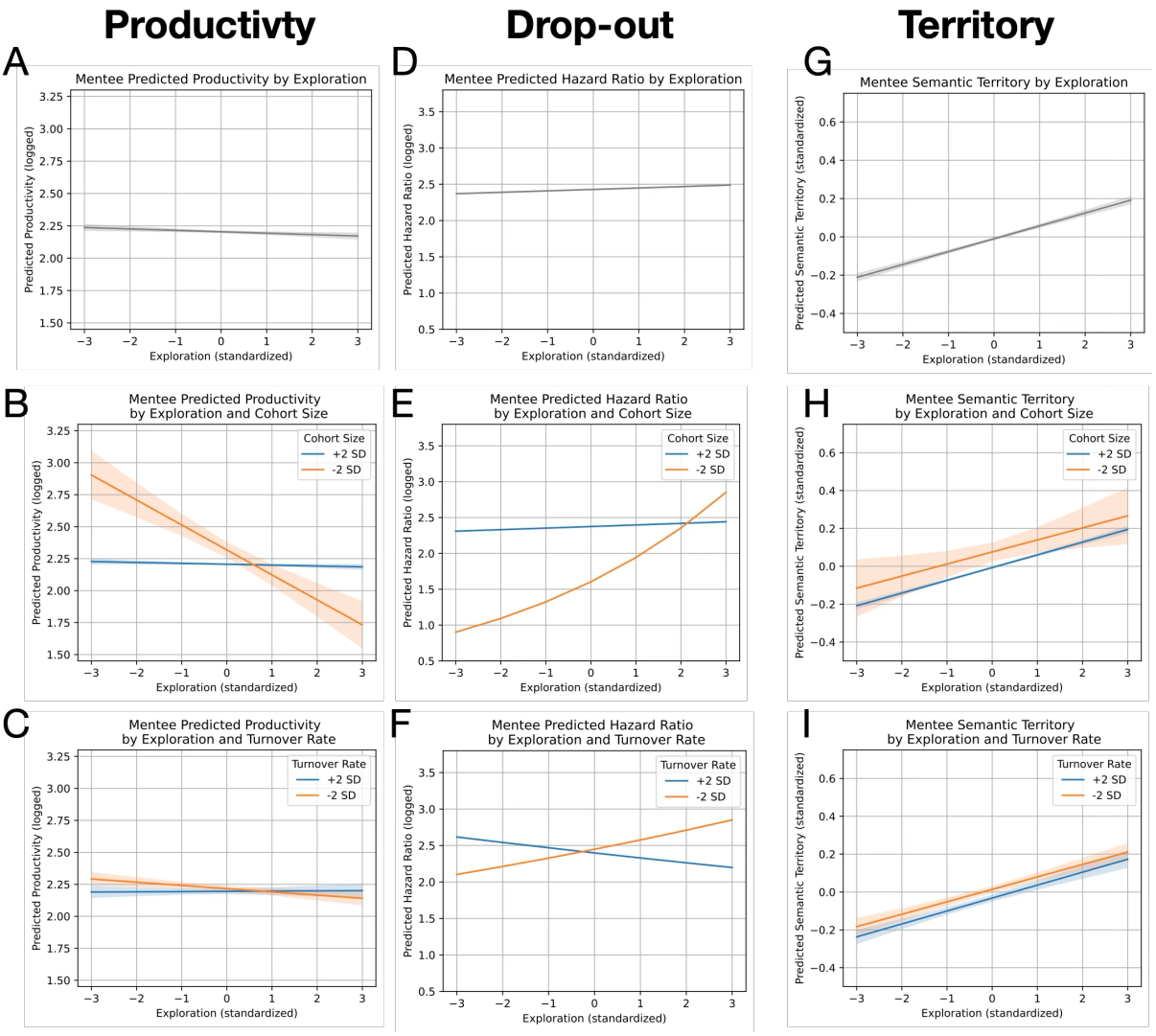

Fig.5 The effects of exploration on mentee-level outcomes. The mentee-level outcomes include total publication productivity (A–C), career hazards (D–F), and semantic territory covered (G–I). Effects are shown at average values (top row) and conditional on cohort size (middle row) and turnover rate (bottom row). Blue lines correspond to predicted outcomes at cohort size or turnover rates that are 2 standard deviations above the mean, while orange lines show predictions 2 standard deviations below the mean. Estimates are model-based predictions from Tables 4–6, with shaded area indicating 90% confidence intervals. For the analysis of publication productivity (A-C), the career longevity is controlled for; and for the analysis of semantic territory (G-I), the publication productivity is controlled for.

Table 1. Mentor Productivity (OLS)

	Dependent variable:		
	Number of Papers		
	(1)	(2)	(3)
exploration_std	-0.236***	-0.226***	-0.237***
	(0.014)	(0.051)	(0.014)
family_size_ln	0.596***	0.597***	0.596***
	(0.023)	(0.022)	(0.023)
turnover_rate_std	-0.048***	-0.048***	-0.048***
	(0.015)	(0.015)	(0.015)
mentor_elite_univ	0.093***	0.093***	0.093***
	(0.023)	(0.023)	(0.023)
mentor_average_LogC10	0.162***	0.162***	0.162***
	(0.017)	(0.017)	(0.017)
exploration_std:family_size_ln		-0.006	
		(0.031)	
exploration_std:turnover_rate_std			0.003
			(0.014)
Constant	-1.657***	-1.656***	-1.661***
	(0.208)	(0.208)	(0.209)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	17,010	17,010	17,010
R^2	0.172	0.172	0.172
Adjusted R^2	0.170	0.170	0.170
Residual Std. Error	1.210 (df = 16956)	1.210 (df = 16955)	1.210 (df = 16955)
F Statistic	66.637*** (df = 53; 16956)	65.401*** (df = 54; 16955)	65.401*** (df = 54; 16955)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 2. Number of Future Mentors (Poisson Regression)

	Dependent variable:		
	Number of Future Supervisors		
	(1)	(2)	(3)
exploration_std	0.0001	-0.033	-0.002
	(0.022)	(0.077)	(0.022)
family_size_ln	1.403***	1.401***	1.402***
	(0.033)	(0.033)	(0.033)
turnover_rate_std	-0.072**	-0.072**	-0.067**
	(0.029)	(0.029)	(0.030)
mentor_elite_univ	0.139***	0.139***	0.140***
	(0.046)	(0.046)	(0.046)
mentor_average_LogC10	0.235***	0.235***	0.234***
	(0.026)	(0.026)	(0.026)
exploration_std:family_size_ln		0.016	
		(0.038)	
exploration_std:turnover_rate_std			0.036*
			(0.021)
Constant	-4.013***	-4.009***	-4.059***
	(0.777)	(0.777)	(0.775)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	17,010	17,010	17,010
Log Likelihood	-7,785.071	-7,784.955	-7,783.535
Akaike Inf. Crit.	15,678.140	15,679.910	15,677.070
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 3. Mentor Semantic Territory (OLS)

	Dependent variable:		
	Semantic Territory		
	(1)	(2)	(3)
exploration_std	0.044***	-0.106***	0.041***
	(0.008)	(0.029)	(0.008)
family_size_ln	0.761***	0.755***	0.761***
	(0.016)	(0.016)	(0.016)
turnover_rate_std	0.083***	0.084***	0.086***
	(0.011)	(0.011)	(0.011)
mentor_elite_univ	0.219***	0.219***	0.219***
	(0.018)	(0.018)	(0.018)
mentor_average_LogC10	0.078***	0.080***	0.078***
	(0.009)	(0.009)	(0.009)
exploration_std:family_size_ln		0.091***	
		(0.017)	
exploration_std:turnover_rate_std			0.019**
			(0.008)
Constant	-1.345***	-1.360***	-1.369***
	(0.168)	(0.167)	(0.169)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	17,010	17,010	17,010
R^2	0.225	0.227	0.225
Adjusted R^2	0.223	0.224	0.223
Residual Std. Error	0.882 (df = 16956)	0.881 (df = 16955)	0.881 (df = 16955)
F Statistic	92.971*** (df = 53; 16956)	92.005*** (df = 54; 16955)	91.412*** (df = 54; 16955)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 4. Mentee Productivity (OLS Random Effects)

	Dependent variable:		
	Number of Papers		
	(1)	(2)	(3)
exploration_std	-0.011**	-0.101***	-0.012**
	(0.005)	(0.020)	(0.005)
family_size_ln	-0.025***	-0.028***	-0.025***
	(0.010)	(0.010)	(0.010)
turnover_rate_std	-0.007	-0.006	-0.005
	(0.007)	(0.006)	(0.007)
career_len	0.143***	0.143***	0.143***
	(0.001)	(0.001)	(0.001)
mentee_elite_univ	0.061***	0.061***	0.060***
	(0.014)	(0.014)	(0.014)
mentor_elite_univ	-0.001	-0.001	-0.001
	(0.010)	(0.010)	(0.010)
mentor_average_LogC10	0.183***	0.183***	0.183***
	(0.007)	(0.007)	(0.007)
exploration_std:family_size_ln		0.047***	
		(0.010)	
exploration_std:turnover_rate_std			0.007
			(0.005)
Constant	-1.470***	-1.445***	-1.471***
	(0.101)	(0.105)	(0.101)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	68,336	68,336	68,336
R^2	0.593	0.593	0.593
Adjusted R^2	0.593	0.593	0.593
F Statistic	96,659.800***	96,740.800***	96,665.030***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5.Mentee' Career Longevity (Cox Proportional Hazard Regression)

	Dependent variable:		
	Exit Hazards		
	(1)	(2)	(3)
exploration_std	0.008	0.101***	0.011*
	(0.006)	(0.022)	(0.006)
family_size_ln	0.095***	0.098***	0.095***
	(0.010)	(0.011)	(0.010)
turnover_rate_std	-0.001	-0.003	-0.005
	(0.008)	(0.008)	(0.008)
mentee_elite_univ	-0.163***	-0.163***	-0.162***
	(0.018)	(0.018)	(0.018)
mentor_elite_univ	-0.058***	-0.058***	-0.059***
	(0.013)	(0.013)	(0.013)
mentor_average_LogC10	-0.187***	-0.188***	-0.188***
	(0.007)	(0.007)	(0.007)
exploration_std:family_size_ln		-0.046***	
		(0.010)	
exploration_std:turnover_rate_std			-0.020***
			(0.005)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	68,336	68,336	68,336
R^2	0.034	0.034	0.034
Max. Possible R^2	1.000	1.000	1.000
Log Likelihood	-413,834.000	-413,824.300	-413,827.400
Wald Test	2,407.890*** (df = 48)	2,424.280*** (df = 49)	2,421.420*** (df = 49)
LR Test	2,364.337*** (df = 48)	2,383.793*** (df = 49)	2,377.581*** (df = 49)
Score (Logrank) Test	2,427.802*** (df = 48)	2,444.051*** (df = 49)	2,441.284*** (df = 49)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6. Mentee Semantic Territory (OLS Random Effects)

	Dependent variable:		
	Semantic Territory		
	(1)	(2)	(3)
exploration_std	0.067***	0.065***	0.067***
	(0.004)	(0.016)	(0.004)
family_size_ln	-0.021***	-0.021***	-0.021***
	(0.007)	(0.007)	(0.007)
turnover_rate_std	-0.012**	-0.012**	-0.011**
	(0.005)	(0.005)	(0.005)
total_pub_ln	0.644***	0.644***	0.644***
	(0.004)	(0.004)	(0.004)
mentee_elite_univ	-0.002	-0.002	-0.002
	(0.010)	(0.010)	(0.010)
mentor_elite_univ	0.007	0.007	0.007
	(0.008)	(0.008)	(0.008)
mentor_average_LogC10	0.008	0.008	0.008
	(0.005)	(0.005)	(0.005)
exploration_std:family_size_ln		0.001	
		(0.008)	
exploration_std:turnover_rate_std			0.001
			(0.004)
Constant	-1.694***	-1.693***	-1.694***
	(0.134)	(0.134)	(0.134)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	51,363	51,363	51,363
R^2	0.556	0.556	0.556
Adjusted R^2	0.555	0.555	0.555
F Statistic	64,222.750***	64,221.120***	64,221.200***
Note:			*p<0.1; **p<0.05; ***p<0.01

References

Bastalich, W. (2017). Content and context in knowledge production: a critical review of doctoral supervision literature. *Studies in Higher Education, 42*(7), 1145-1157. doi:10.1080/03075079.2015.1079702

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy, 33*(4), 599-616.

BROWN, G., & ATKINS, M. (1988). *Effective Teaching in Higher Education*. London: Methuen.

Cheng, M., Smith, D. S., Ren, X., Cao, H., Smith, S., & McFarland, D. A. (2023). How New Ideas Diffuse in Science. *American Sociological Review, 88*(3), 522-561. doi:10.1177/00031224231166955

Damaceno, R. J. P., Rossi, L., Mugnaini, R., & Mena-Chalco, J. P. (2019). The Brazilian academic genealogy: evidence of advisor-advisee relationships through quantitative analysis. *Scientometrics, 119*(1), 303-333. doi:10.1007/s11192-019-03023-0

David, S. V., & Hayden, B. Y. (2012). Neurotree: A Collaborative, Graphical Database of the Academic Genealogy of Neuroscience. *Plos One, 7*(10). doi:10.1371/journal.pone.0046608

Dores, W., Benevenuto, F., Laender, A. H. F., & Ieee. (2016). *Extracting Academic Genealogy Trees from the Networked Digital Library of Theses and Dissertations*. New York: Ieee.

Hamilton, W. D., & May, R. M. (1977). DISPERSAL IN STABLE HABITATS. *Nature, 269*(5629), 578-581. doi:10.1038/269578a0

Hockey, J. (1991). THE SOCIAL-SCIENCE PHD - A LITERATURE-REVIEW. *Studies in Higher Education, 16*(3), 319-332. doi:10.1080/03075079112331382875

Hockey, J. (1996). Motives and meaning amongst PhD supervisors in the social sciences. *British Journal of Sociology of Education, 17*(4), 489-506. doi:10.1080/0142569960170405

Laudel, G., & Gläser, J. (2008). From apprentice to colleague: The metamorphosis of Early Career Researchers. *Higher Education, 55*(3), 387-406. doi:10.1007/s10734-007-9063-7

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. doi:10.1038/s41597-023-02198-9

Malmgren, R. D., Ottino, J. M., & Amaral, L. A. N. (2010). The role of mentorship in protege performance. *Nature, 465*(7298), 622-U117. doi:10.1038/nature09040

Marsh, H. W., Rowe, K. J., & Martin, A. (2002). PhD students' evaluations of research supervision - Issues, complexities, and challenges in a nationwide Australian experiment in benchmarking universities. *Journal of Higher Education, 73*(3), 313-348. doi:10.1353/jhe.2002.0028

Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics, 9*(4), 962-973. doi:https://doi.org/10.1016/j.joi.2015.10.005

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rossi, L., Damaceno, R. J. P., Freire, I. L., Bechara, E. J. H., & Mena-Chalco, J. P. (2018). Topological metrics in academic genealogy graphs. *Journal of Informetrics, 12*(4), 1042-1058. doi:10.1016/j.joi.2018.08.004

Rossi, L., Freire, I. L., & Mena-Chalco, J. P. (2017). Genealogical index: A metric to analyze advisor-advisee relationships. *Journal of Informetrics, 11*(2), 564-582. doi:10.1016/j.joi.2017.04.001

Shibayama, S. (2019). Sustainable development of science and scientists: Academic training in life science labs. *Research Policy, 48*(3), 676-692. doi:10.1016/j.respol.2018.10.030

Shibayama, S., Baba, Y., & Walsh, J. P. (2015). Organizational Design of University Laboratories: Task Allocation and Lab Performance in Japanese Bioscience Laboratories. *Research Policy, 44*(3), 610–622.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science, 354*(6312), aaf5239.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Stephan, P. E. (2012). *How economics shapes science*. Cambridge, MA: Harvard University Press.

Wang, J., & Shibayama, S. (2022). Mentorship and creativity: Effects of mentor creativity and mentoring style. *Research Policy, 51*(3), 104451. doi:https://doi.org/10.1016/j.respol.2021.104451

Appendix

Tables

Table A1 Matching Mentors and Mentees from PTDG Global data

Criterion	Condition / Rule	Score	Rationale
Middle Name Match	Exact match (e.g., *John Alan Smith* ↔ *John Alan Smith*)	3	Strong indicator of identity
	Partial match (e.g., *John A. Smith* ↔ *John Alan Smith*)	2	High likelihood of same person
	One middle name missing (e.g., *John A. Smith* ↔ *John Smith*)	1	Still a plausible match
	Conflicting middle names (e.g., *John Alan Smith* ↔ *John B. Smith*)	–2	Likely different individuals
Name Rarity	Name block appears ≤ 5 times	1	Rare names are more reliably matched
	Name block appears 50 times (top 5% of distribution)	–1	Common names are more prone to false positives
Temporal Plausibility (Gestation Time)	8–25 years between advisor and student PhDs	2	Typical academic progression
	5–8 years	1	Plausible, though fast progression
	<5 years or >25 years	–2	Unlikely supervisor–student relationship
Field Similarity	Cosine similarity ≥ 0.95	2	Very strong topical alignment
	Cosine similarity ≥ 0.90 and < 0.95	1	Moderate topical alignment
	Cosine similarity < 0.90	–2	Weak alignment, less likely relationship

Table A2 Matching PQDT Global Authors to SciSciNet Authors.

Criterion	Condition / Rule	Score	Rationale
Middle Name Match	Exact / Partial / One Missing / Mismatch	+3 / +2 / +1 / –2	More precise matches receive higher scores.
Year Difference	$\Delta_{year} \leq 5$	+2	Dissertation and Publication should occur within comparable time window.
Title Similarity (SPECTER2)	≥ 0.98 / ≥ 0.90 / < 0.72 (median) / else	+3 / +2 / –1 / 0	Cosine similarity of title embeddings
Field Similarity	≥ 0.98 / ≥ 0.95 / < 0.95	+2 / +1 / –1	Based on field-level embedding cosine sim
Institution Similarity	≥ 0.98	+2	Same institution matches receive higher score.
University Affiliation	True	+1	University affiliation of publication receives higher score.
Name Rarity	≤ 5 / ≥ 50 / else	+2 / –1 / 0	Based on first–last block frequency

Figures

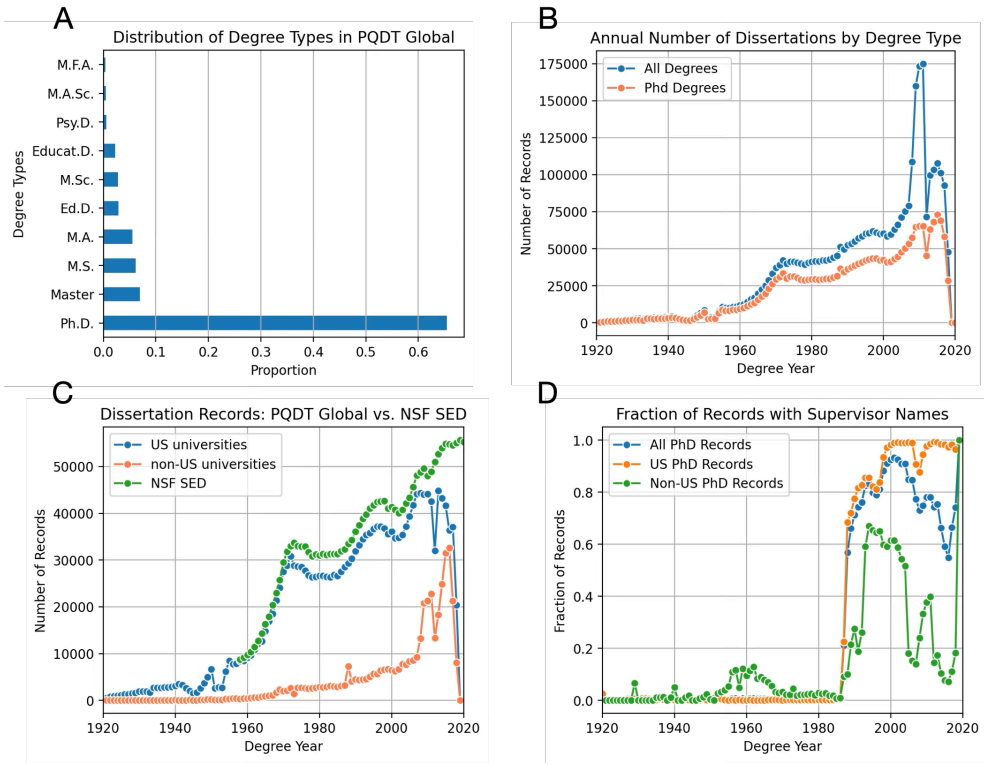

Fig. A1. Coverage and metadata availability in PQDT Global. Panels A–D summarize the distribution, temporal coverage, and supervisor coverage of dissertation records in ProQuest Dissertations & Theses Global (PQDT Global).

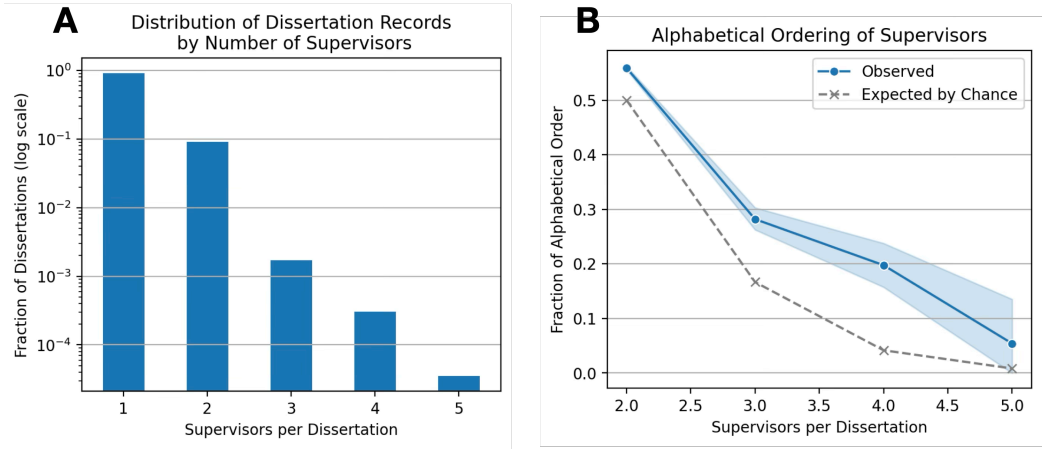

Fig. A2. Supervisor counts and ordering in dissertation records. (A) Distribution of dissertation records by the number of supervisors per dissertation. (B) Fraction of dissertation records in which supervisors are listed in alphabetical order, by the number of supervisors per dissertation. The dashed line indicates the fraction expected under random ordering; shaded bands denote 95% confidence intervals for the observed fractions.

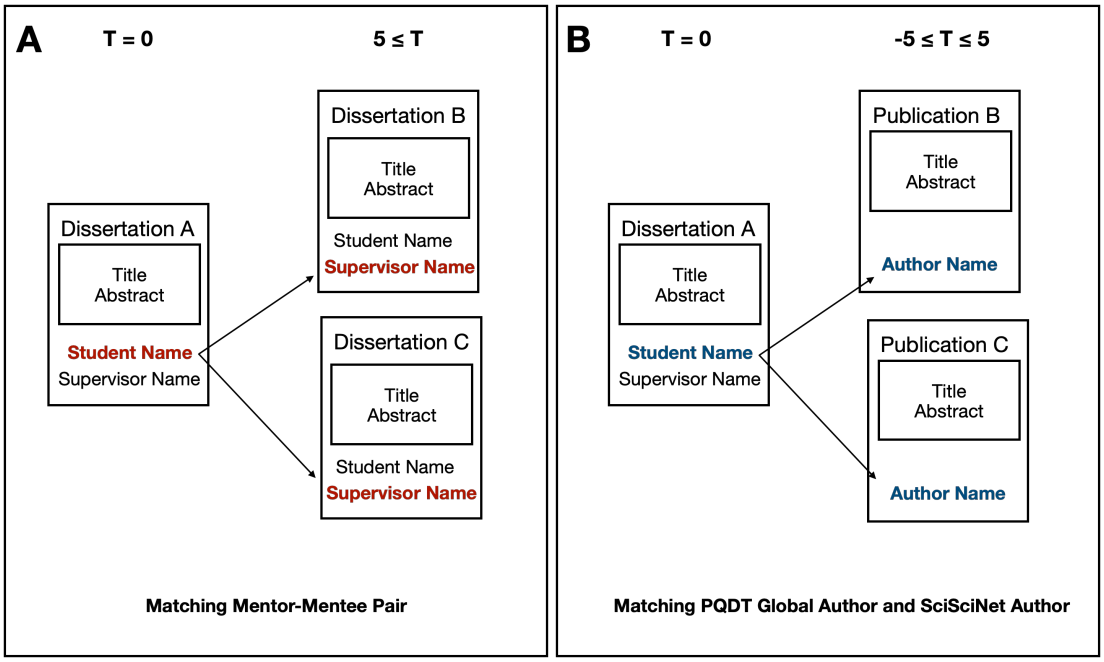

Fig. A3. (A) Matching Mentor-Mentee pairs from PQDT Global and (B) connecting PQDT Global authors with SciSciNet Authors.

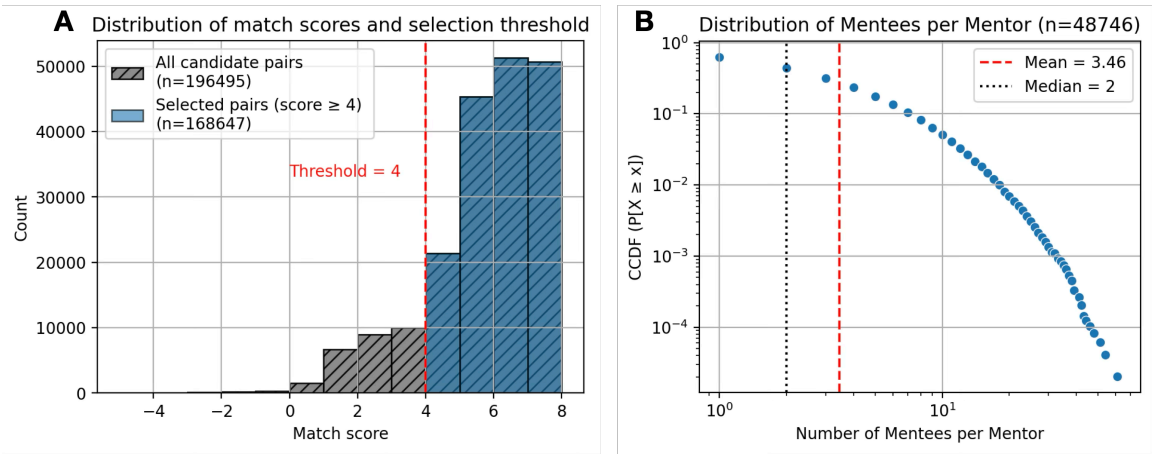

Fig. A4. (A) Distribution of mentor-mentee matched-scores and selection threshold (B) Distribution of Mentees per Mentor

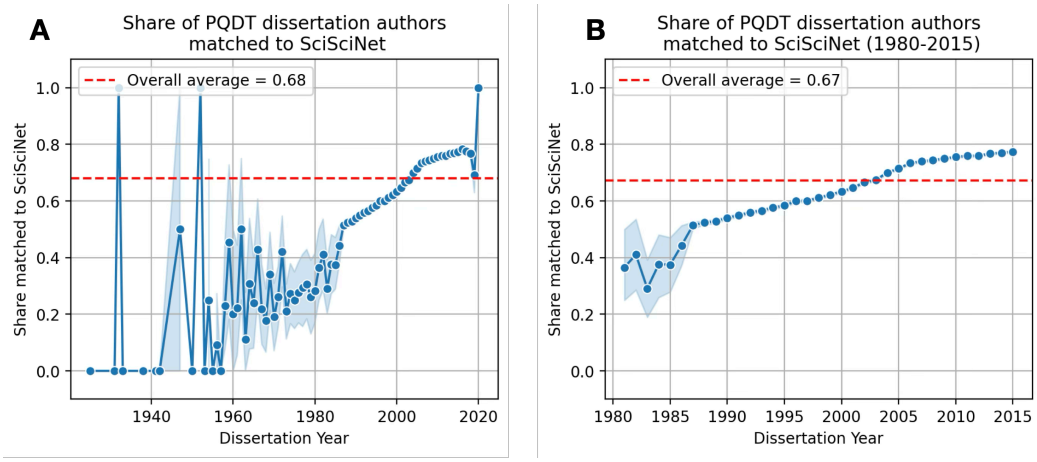

Fig. A5. (A) Share of PQDT Global Dissertation authors that are matched to SciSciNet authors. (B) Selected Share (from 1980 to 2015) of PQDT Global Dissertation authors that are matched to SciSciNet authors.

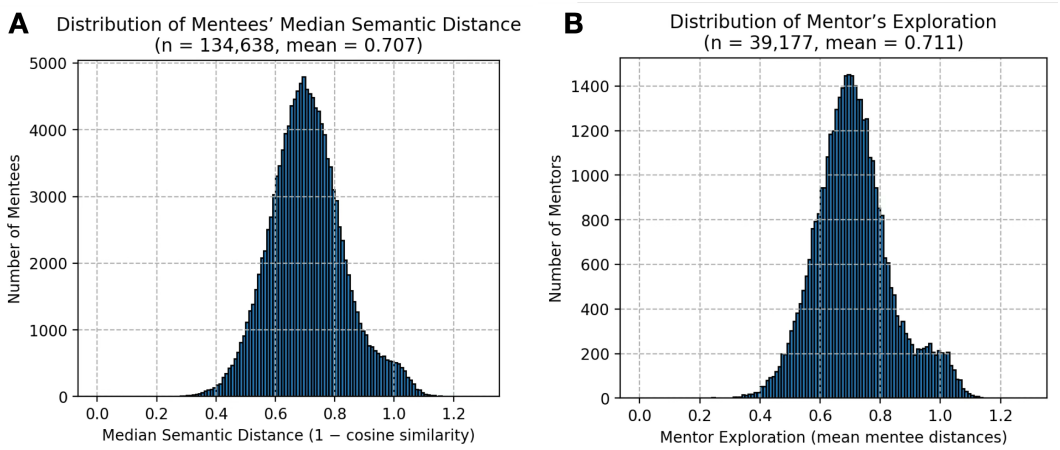

Fig. A6. (A) Distribution of Mentee's Median Semantic Distance. (B) Distribution of Mentor's Exploration.

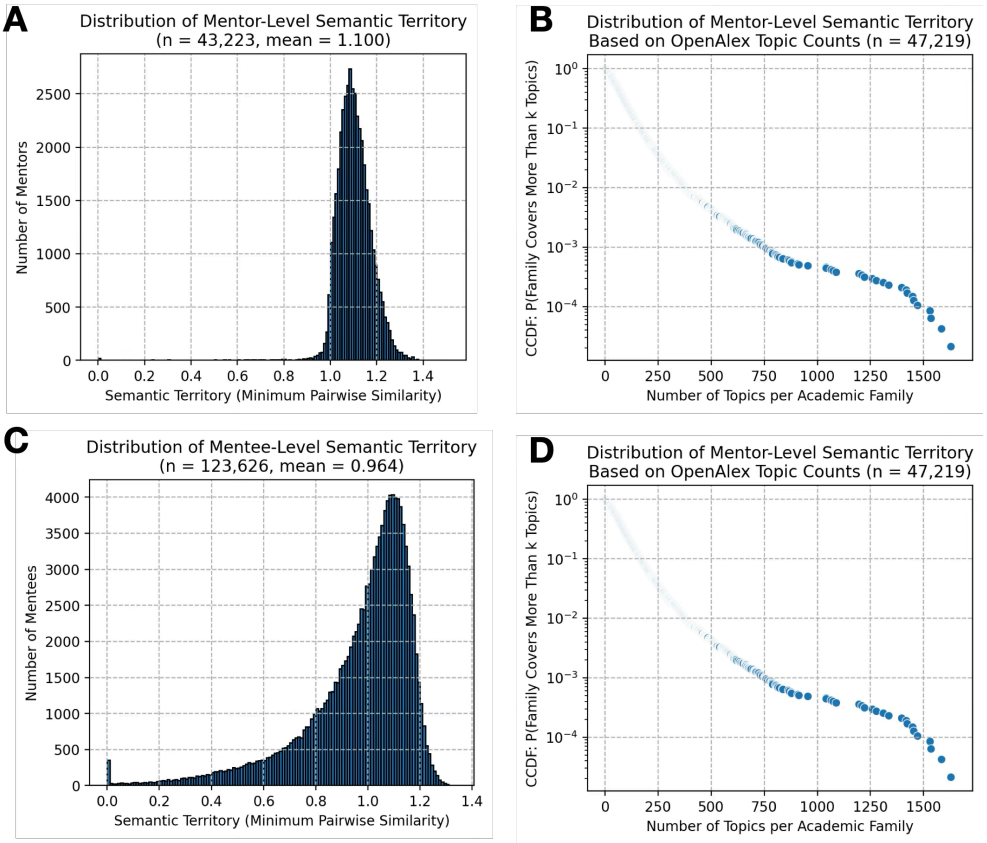

Fig. A7. (A) Distribution of Mentor-Level Semantic Territory. (B) Distribution of Mentor-level Semantic Territory Based on OpenAlex Topic Counts. (C) Distribution of Mentee-Level Semantic Territory.