REPRODUCTION OF SCIENTISTS

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

[a] KAIST, Graduate School of Science and Technology Policy, Daejeon, South Korea

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

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

Abstract

Academic family lineages, formed through chains of mentor–mentee relationships, provide the critical pathway through which scientists are trained and scientific knowledge is transmitted. These lineages are subject to competition, which can expand, shrink, and terminate the lineages. Such evolutionary fates are shaped by scientists' strategic choices in navigating the scientific space. Doctoral training represents a pivotal juncture in this evolutionary process. Here we investigate the extent to which mentees' research topics diverge from their mentors'—conceptualized as exploration—and how exploration affects the family evolution in various conditions, drawing on a novel dataset linking 17,010 U.S. PhD mentors and 78,102 mentees to large-scale bibliometric records. Our results are twofold. First, we identify tensions among family-level goals: exploration reduces mentors' productivity but expands family semantic territory, with this trade-off especially pronounced in large families and fast-growing fields. In contrast, refraining from exploration is optimal across the goals in small families and established fields. Second, our mentee-level analyses reveal the underlying mechanisms. Exploration compromises individual mentees' productivity and career longevity on average, yet leaves a lasting impact on their topic choice, contributing to families' territorial expansion, particularly in fast-growing fields. In small families, however, the negative effects dominate, leading exploring families to lose semantic territories.

Keywords: Academic Genealogy, Doctoral Training, Sociology of Science

Introduction

As science underpins our modern knowledge-based society, it is essential that its foundational input—scientists—be sustainably reproduced (*1-3*). 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 (*2, 4-7*).

Importantly, academic lineages are subject to competition – failing to win the competition can shrink the line of research, give space away to other lineages, and potentially terminate the lineage. 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 supervision of their mentors, which influences the trajectories of their research in the long term (*8*). 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 physical distance between the sites of parents and offspring—is a key strategic factor (*9*), and 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 fate of academic lineages. Here, a stereotypical assumption is that mentees usually follow their mentors' research agendas with limited exploration (*10*), but closer examination reveals variation across mentors and mentees (*11*). In this study, we inquire how this varying degree of exploration affects an academic family's long-term outcomes such as productivity, survival, and semantic territory, and how it is conditioned on 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 (*4-6, 12*). While the literature illustrates the topological features of lineage evolution and their temporal dynamics (*4, 5, 13, 14*), 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 (*7, 15*). While the literature has revealed substantial heterogeneity in the practices in the postgraduate training context (*16-19*), it has paid limited attention to the consequences of those practices, especially at the academic family level. This study aims to bridge these gaps.

To this end, we draw on a newly constructed dataset that links U.S. doctoral dissertation records to large-scale bibliometric data of 17,010 mentors and 78,102 mentees. The result of our analyses is twofold. First, it highlights the incompatibility of 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 is 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, illustrating how exploration during PhD training affects the mentee-level outcomes, which collectively shape family-level goals.

Results

We constructed data of academic lineages based on PhD mentor-mentee relationships, drawing on the database of dissertation in 1991-2015 (Fig.1A). We identified 17,010 mentors who supervised at least three mentees. Then, we define a family as a group of scientists consisting of a mentor and his/her all mentees throughout the mentor's career (Fig.1B). The mean family size is 7.41 (Fig.1C).

We then examine the family's exploration strategy, which we define as the degree to which mentors allow their mentees to engage in research topics remote from their own. This measure is operationalized as the semantic distance between mentees' dissertations and mentors' publications, by measuring the cosine distance (1 – cosine similarity) between the document embeddings (Fig.2A). The mean of exploration is 0.70, suggesting reasonable similarity between the pairs (Fig.2B).

We investigate how such varying levels of exploration affect family-level outcomes that mentors may be interested in pursuing (Fig.3). For one, scientists in a mentorship role tend to be under competition, which can make them prioritize their own productivity, potentially at the cost of their mentees' (*7*). On the other hand, mentors may consider doctoral education as an obligation to the academic community (*20*), 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. In what follows, we examine how exploration affects these family-level goals: (1) the mentor's own productivity (irrespective of the fate of the family) (Fig.3A), (2) the survival of the family lineage (Fig.3B), and (3) the expansion of the semantic territory of the family (Fig.3C). Note that these goals are interrelated and can be in conflict.

We further probe the effect of exploration from two angles. First, we analyze how different contexts change the effect of exploration. In particular, we highlight two contextual factors – 1) family size, or the number of mentees supervised by a mentor (Fig.1C), and 2) field dynamics, or the rate at which research topics in a field are replaced (Fig.4) – which are likely to change the rationale of exploratory strategy. Second, on top of family-level outcomes, we also analyze outcomes at the individual mentee's level to elucidate the underlying mechanisms.

Exploration and family evolution. First, we analyze the relationship between exploration and three family-level goals by regression analyses, in which we control for several confounding factors as well as test the moderating effects by family size and field dynamics (Table S1-S3 in SI). Based on the regression analyses, Fig.5 visualizes the predicted relationships between exploration and the family-level outcomes: the mentors' publication productivity through their career (Fig.5A), the number of future mentors trained (Fig.5B), and the semantic territory covered by the family members (Fig.5C). The top row (Fig.5A-C) concerns the direct effect of

exploration, while the middle row (Fig.5D-F) and the bottom row (Fig.5G-I) illustrate the moderating effects by family size and field dynamics, respectively.

First, Fig.5A shows that greater exploration is negatively associated with mentors' publication productivity, suggesting that mentors who allow mentees to pursue topics remote from their own tend to publish less than those who direct mentees to remain closely aligned. Fig.5D illustrates the moderating effect of family size by predicting the mentor's productivity for small families (orange) and for large families (blue). It shows that mentor's productivity is overall higher in larger families, and the parallel lines suggest that family size does not influence the disadvantages caused by exploration. Similarly, Fig.5G examines the moderating effect of field dynamics, where the mentor's productivity is predicted for fast-evolving fields (blue) and slow-evolving fields (orange), suggesting that the negative relationship between exploration and the mentor's productivity is also not affected by field dynamics.

Regarding the continuation of the academic lineage, Fig.5B indicates that the number of future mentors is not affected by exploration, on average. Examining the moderating effect of family size, Fig.5E shows that this null relationship holds for both large and small families. However, we find that field dynamics condition this relationship (Fig.5H). Overall, families operating in fast-evolving fields (blue) generate fewer future mentors than those operating in slow-evolving fields (orange), but the former gain from exploration while the latter lose by exploration. Thus, in rapidly developing fields, allowing mentees to explore distant topics is associated with a greater chance of lineage persistence.

As to the semantic territory, we quantify the breadth of research topics covered by the family, based on text embeddings and topic classifications (Section S.7 in SI). Fig.5C shows that exploration is, on average, positively associated with greater semantic territory. In other words, families in which mentees are allowed to deviate from their mentor's topic tend to cover a broader semantic territory in the long term. This relationship, however, differs by family size (Fig.5F). Exploration is associated with greater semantic territory in large families (blue), whereas it is associated with narrower territory in small families. Field dynamics also condition the relationship (Fig.5I) – exploration is more advantageous in fast-evolving fields (blue) to cover broader semantic territories.

These results highlight trade-offs between the goals in determining the degree of exploration. On average, exploration appears detrimental to mentors' productivity but preferable for families' semantic territory. The analyses of moderating effects suggest that this trade-off becomes pronounced in large families as well as in fast-growing 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.

Exploration and mentee outcomes. To elucidate the underlying mechanisms, we examine the relationship between exploration and the three mentee-level outcomes: (1) individual mentees'

publication productivity, (2) career dropout rates, and (3) semantic coverage, respectively corresponding to the family-level outcomes. We similarly run regression analyses (Table S4-S6 in SI) and illustrate the predicted relationships between exploration and the mentee-level outcomes (Fig.6).

First, we analyze the relationship between exploration and mentees' career longevity by survival analysis (Fig.6B). Fig.6B suggests that, on average, exploration is associated with slightly higher hazard rates of career exit, which is operationalized as termination of scholarly publication. In other words, exploration tends to result in shorter career longevity. This suggests that it is more challenging for mentees to find jobs in research areas remote from mentors' than in proximity, perhaps because mentees cannot exploit mentors' prestige in recruitment. This effect is conditioned on family size (Fig.6E). On average, mentees from larger families (blue) face higher career exit hazards than those from smaller families (orange). This implies that members in large families may compete with one another for job opportunities. Fig.6E also shows that the effect of exploration is negative only in smaller families (orange), while exploration is desirable in larger families. Thus, the advantage in finding job opportunities in mentors' proximity is substantiated only if the family is small, and taking distance from mentors may help avoid competition among family members. Further, Fig.6H indicates that mentees in fast-evolving fields (blue) experience a negative association between exploration and the career exit hazard, whereas those in slow-evolving fields (orange) experience a positive association. In other words, exploration is a favorable strategy for mentees trained in fast-evolving fields, while it is the opposite for mentees in slow-evolving fields. This may be because exploration can help mentees find job opportunities emerging in fast-growing fields. At the family level, this is manifested as a greater number of future mentors from families in fast-growing fields (Fig.5H).

Next, we analyze the relationship between exploration and mentees' publication productivity after controlling for career longevity. Fig.6A shows a negative association between exploration and mentees' publication productivity on average. This suggests the inefficiency of exploration in training – that is, mentees fail to utilize mentors' expertise. At the family level, this negative effect is directly translated into lower productivity of exploring mentors who may be penalized by limited mentor-mentee collaboration (Fig.5A). We then analyze the moderating effect by family size. Fig.6D first shows that mentees in small families (orange) gain productivity by not exploring, implying that mentees can exploit mentors' expertise effectively. This gain from non-exploration, i.e., through tight mentor-mentee relationships, is directly translated into mentors' productivity (Fig.5D). In contrast, in large families, exploration is favorable (blue). This is perhaps because mentees in large families have to compete one another for mentors' capacity, and this negative effect by competition appears mitigated by exploration. This gain from exploration, i.e., weak mentor-mentee relationships, does not contribute to mentor's productivity (Fig.5D). Fig.6G then shows that exploration is detrimental to mentees' productivity only in slow-evolving fields (orange). In fast-evolving fields (blue), the relationship between exploration and mentees' productivity is almost flat. Thus, exploration may help gain from opportunities emerging in dynamic fields, compensating the disadvantage of exploration.

Lastly, we examine the effect of exploration on individual mentees' semantic territory with controlling for their publication productivity. Fig.6C shows that, on average, the exploration strategy is positively associated with the semantic territory covered by mentees. In other words, mentees trained under academic families whose dissertation topics deviate from their mentors' research areas tend to cover a broader semantic territory over the course of their careers than those who remain closely aligned with their mentors. This implies that exploration during the PhD period has a long-lasting impact on mentees' broader topic choice in later careers (*8*). The positive effect of exploration does not seem affected by family size (Fig.6F) or field dynamics (Fig.6I). At the family level, the positive effect of exploration is translated into the wider semantic territory of exploring families (Fig.5C). We observed that this family-level positive effect is reduced in slow-evolving fields (Fig.5I) and reversed in small families (Fig.5F), possibly because of the aforementioned disadvantages of exploration for career longevity under such conditions (Fig.6E and 6H).

Conclusions

Mentor-mentee relationships form academic family lineages, offering a critical path through which scientists are reproduced (*2, 4-7*). This study focuses on the exploration of mentees' topics as mentors' key strategy and inquires how it shapes the evolution of academic families. Our analyses highlight incompatibility among the 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, but non-exploration is desirable in other conditions.

The mentee-level analyses help elucidate underlying mechanisms. Namely, exploration during PhD training is on average inefficient for individual mentees' productivity, which is translated into mentors' lower productivity perhaps through the channel of mentor-mentee collaboration. Exploration however makes a lasting impact on mentees' topic selection, leading to a greater semantic coverage of the family. Exploration also decreases individual mentees' career longevity on average, but exploration is favorable in fast-growing fields, which appears to help grow family size and semantic territory in such fields. Finally, in small families, refraining from exploration increases individual mentees' productivity as well as career longevity. This effect is substantial – when aggregated to the family level, small families refraining from exploration ultimately achieve greater semantic territory than those that do explore, in contrast to the positive effect of exploration observed in large families.

Our results offer a few policy implications. For example, incentives for exploration and novelty may disadvantage small families, where the cost of exploration is less easily absorbed and can ultimately undermine family-level knowledge development. In contrast, exploration appears more beneficial in large families and fast-growing fields, where individual mentees' risks are more likely to translate into durable expansions of family territory. This suggests that policies on doctoral education (funding, etc.) should adopt a context-sensitive design, differentiating support for exploration according to the organization of research communities (family size)

and the dynamics of fields, and explicitly balancing individual mentees' career sustainability with community-level knowledge development.

Future research can extend this work in a few directions. First, while this study focuses on topic exploration, 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 us more directly to 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.

Materials and Methods

Data. We integrate large-scale doctoral training records with bibliometric data to examine how mentoring strategies shape the long-run academic outcomes of both mentor and mentees. We draw mentor-mentee relationships from ProQuest Dissertations and Theses Global, which provides metadata on dissertation authors, supervisors, degree years, fields, and titles and abstracts. We restrict the sample to PhD dissertations granted by US institutions to ensure consistent coverage of the mentor-mentee relationship. Publication records for both mentor and mentees are then obtained from SciSciNet (*21*), a large-scale bibliometric data lake built on the Microsoft Academic Graph (Section S1 in SI).

Construction of Academic Lineages. We identify academic genealogies by tracing individuals who first appear as dissertation authors and later reappear as main supervisors on subsequent dissertations. Each mentee is linked to a single primary mentor, while mentors may supervise multiple mentees. Using a multi-step identity-matching procedure, we link dissertation authors to a disambiguated author identifier (MAGID) in SciSciNet, yielding longitudinal publication histories of both mentor and mentees. The final sample consists of 17,010 mentors and 78,102 mentees, spanning multiple cohorts and scientific fields (Section S4 in SI).

Main variables. We focus on three variables associated with the mentee's training environment. First, exploration captures the extent to which mentors allow mentees to pursue research topics that deviate from the mentor's prior work. This measure is operationalized by computing the semantic distance between a mentee's dissertation title and abstracts and the mentor's pre-existing publication title and abstracts using BERT embeddings, and then averaging this distance across all mentees supervised by a mentor. Higher values indicate greater allowance for topic exploration (Fig.2; Section S5 in SI). Second, family size is measured as the number of PhD students supervised by a mentor plus the mentor themselves (Fig.1.B-C). Third, to operationalize field dynamics, we measure the extent to which the topical contents of a scientific field evolve slowly or rapidly over time. In fast-moving fields, newly published work draws on a rapidly changing conceptual vocabulary, whereas in slow-moving fields, the core

topics remain relatively stable. This measure is first constructed at the subfield level based on the average concept overlap between periods. We then map subfield-level dynamics to individual mentors based on the subfields of their publications. Because publications may be associated with multiple fields, we compute a weighted average across fields to obtain a mentor-level measure of field dynamics. This mentor-level field dynamics variable is treated as time-invariant in our analyses (Fig.4; Section S.6 in SI).

We measure three family-level outcomes (Fig.3; Section S8.1 in SI). First, mentors' publication productivity is measured by their total publication count throughout their career. Second, the survival of the family lineage is captured by the number of mentees who later become mentors themselves. Third, the family's semantic territory measures the breadth of research topics covered by mentors' academic families over time, based on text embeddings and topic classifications. We also measure three mentee-level outcome variables related to the family-level goals (Section S8.2 in SI). First, we measure the publication productivity of individual mentees by their total publication count throughout their career. Second, we measure the career longevity of individual mentees by the period in which they publish academic papers. Third, we measure the semantic territory that individual mentees cover through their careers.

Statistical Analysis. We analyze how exploration influences the family-level goals and how these relationships are moderated by family size and field dynamics using regression models:

$$E[Y_j|X_j] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + \beta_4 \text{Exploration}_j \times \text{FamilySize}_j + \beta_5 \text{Exploration}_j \times \text{FieldDynamics}_j + Z_j'\delta + \varepsilon_j \quad (1)$$

Y_j denotes the outcome of interest for family j. Our key explanatory variable are exploration ($Exploration_j$), family size ($FamilySize_j$), and field dynamics ($FieldDynamics_j$). Z_j' is a vector of control variables, including mentor impact, institutional prestige, cohort fixed effects, and field fixed effects (Section S8.1 in SI).

A similar specification is employed to predict the mentee-level goals:

$$E[Y_{ij}|X_{i,j}] = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + \beta_4 \text{Exploration}_j \times \text{FamilySize}_j + \beta_5 \text{Exploration}_j \times \text{FieldDynamics}_j + Z_{i,j}'\delta + \mu_j + \varepsilon_{ij} \quad (2)$$

, where Y_{ij} is the outcome of interest for mentee i trained by mentor j, and μ_j is the mentor-level random effects (Section S8.2 in SI).

We use estimation strategies suitable for each dependent variable. Namely, productivity and semantic territory outcome variables, at both family and mentee levels, are estimated using ordinary least squares regressions. The number of future mentors in academic lineage is estimated using a Poisson regression, and the career exit hazards of mentees are estimated using Cox proportional hazards regression.

Figures

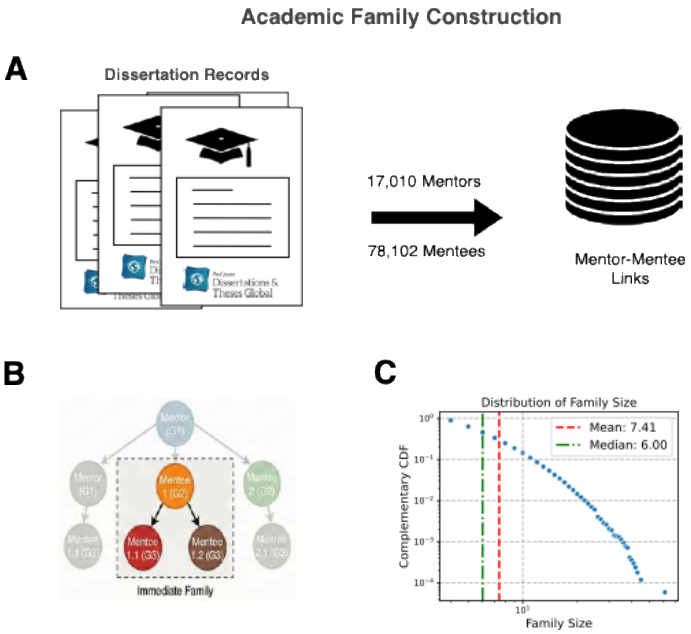

Fig.1. Construction of academic families from dissertation records. Academic family construction. (A) Mentor–mentee pairs are constructed from advisor information in the ProQuest Dissertations & Theses Global database. (B) An academic family is defined as the immediate family surrounding a focal mentor within the mentor–mentee network, including direct doctoral descendants. (C) Distribution of academic family size across mentors in the sample.

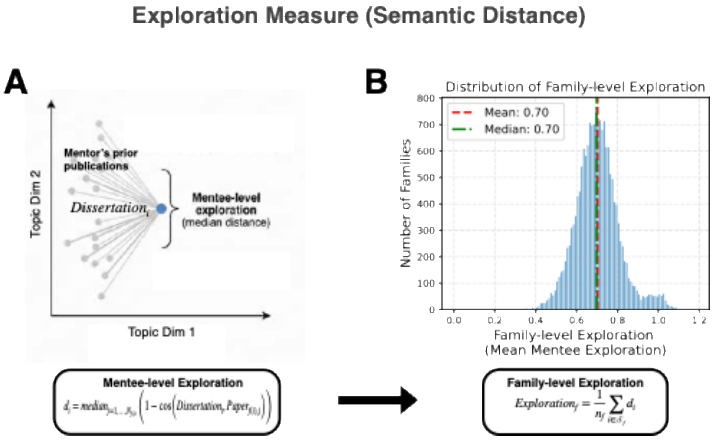

Fig.2. Operationalization of exploration as semantic distance. (A) Mentee-level exploration is measured as the cosine distance between a doctoral dissertation and the mentor's prior publication record in semantic topic space. (B) Family-level exploration is computed as the mean of mentee-level exploration across all mentees within an academic family; the panel shows the distribution of this measure.

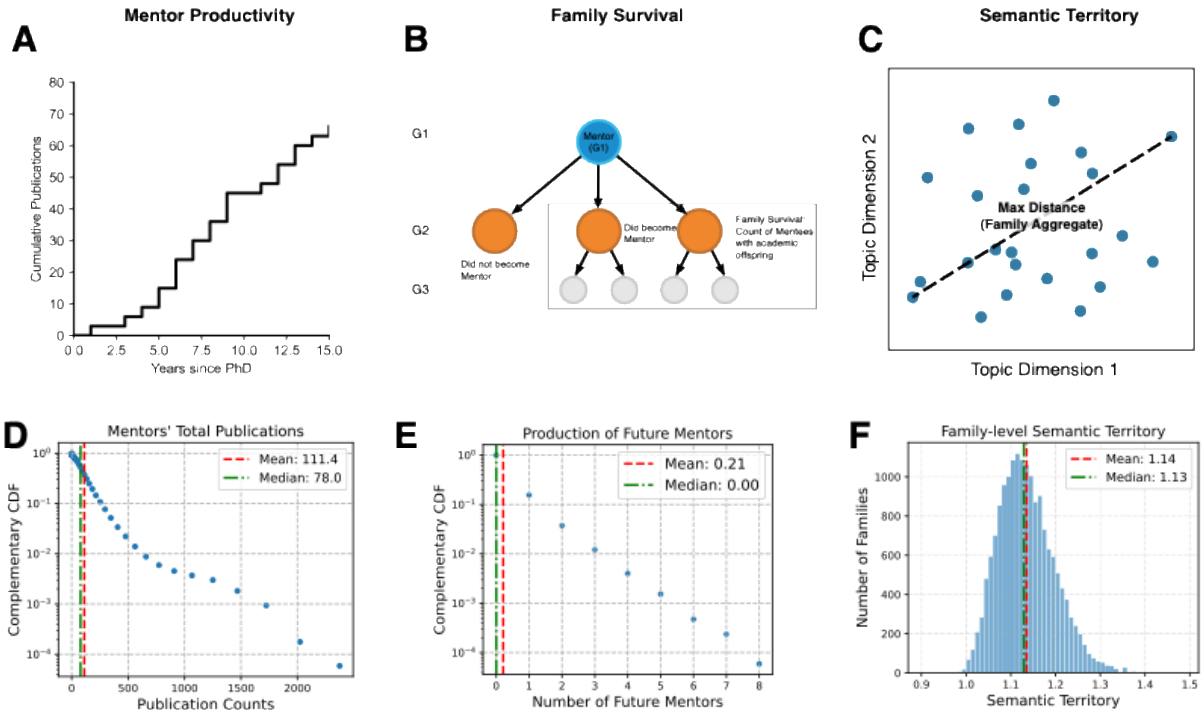

Fig.3. Outcome variables capturing academic family evolution. Panels illustrate the definitions and distributions of the main outcome variables used in the regression analyses. (A) Mentors' total career productivity, measured as the cumulative number of publications over their academic careers. (B) Family survival, measured as the number of a mentor's mentees who later become mentors themselves. (C) Semantic territory, defined as the maximum semantic distance among family-level publications, capturing the breadth of topics covered by a mentor's academic family. (D–F) Empirical distributions of total career productivity, family survival, and semantic territory, respectively. Vertical lines indicate mean and median values where shown.

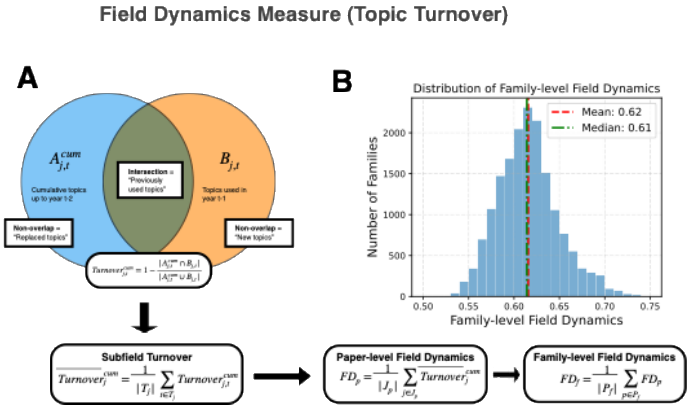

Fig.4. Operationalization of field dynamics as turnover of scientific concepts. (A) Field dynamics are measured as the turnover rate of scientific concepts, operationalized using the proportion of nonoverlapping noun phrases between consecutive time periods. (B) Family-level field dynamics are computed by averaging paper-level turnover across all papers associated with a mentor; the panel shows the distribution of this measure.

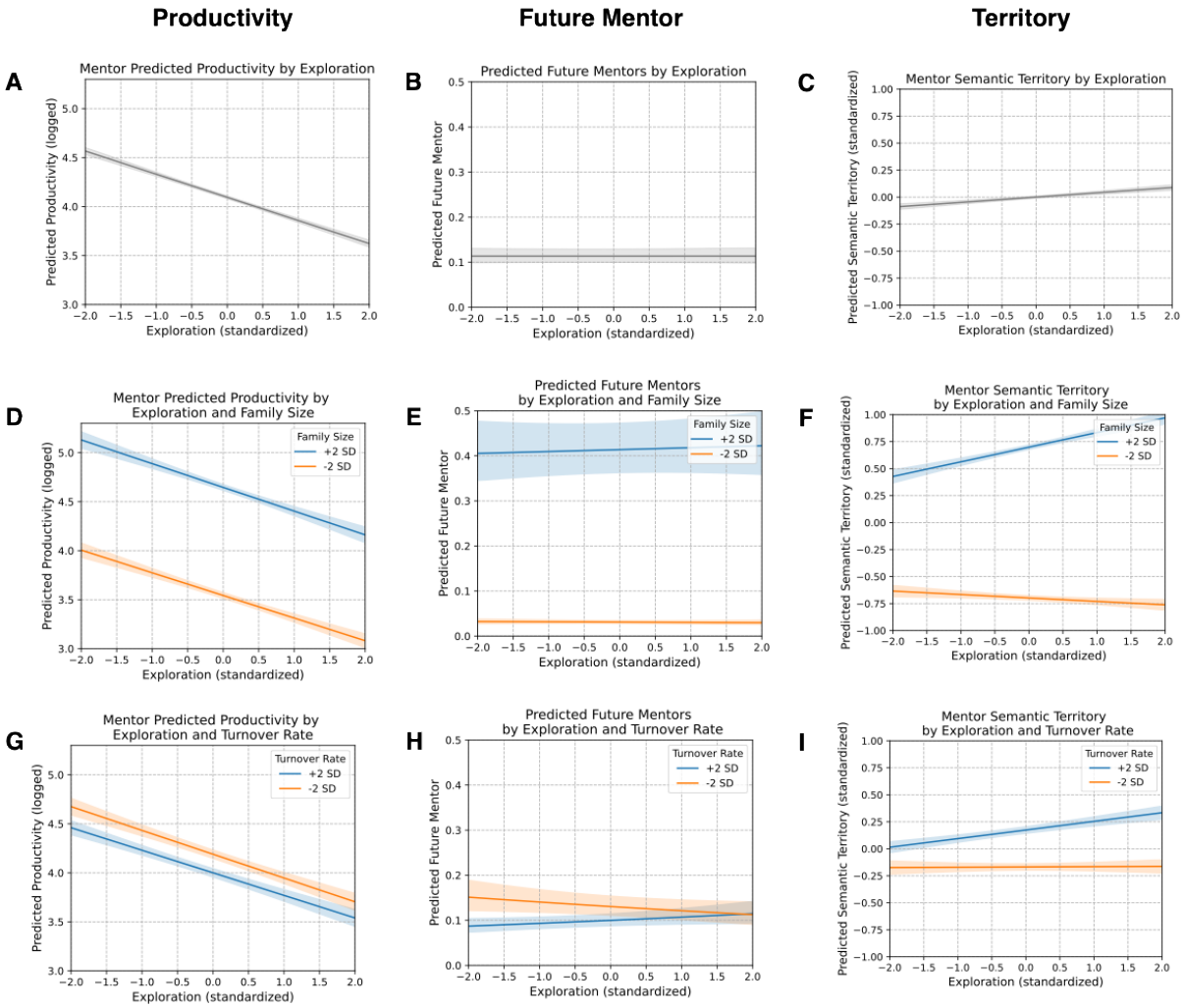

Fig.5. Exploration and Family Evolution Outcomes. Panels show model-predicted outcomes as a function of exploration (standardized). (A,D,G) The main effects of exploration on mentor productivity (log publications), the number of future mentors, and the family's semantic territory, respectively. (B,E,H) Predicted outcomes by exploration interacted with family size (±2 SD). (C,F,I) Predicted outcomes by exploration interacted with field dynamics, measured as conceptual turnover (±2 SD). Shaded bands indicate 90% confidence intervals. All predictions are based on the full multivariate models reported in the Table S1-S3 in SI, with continuous covariates held at their means and indicator variables set to their reference categories.

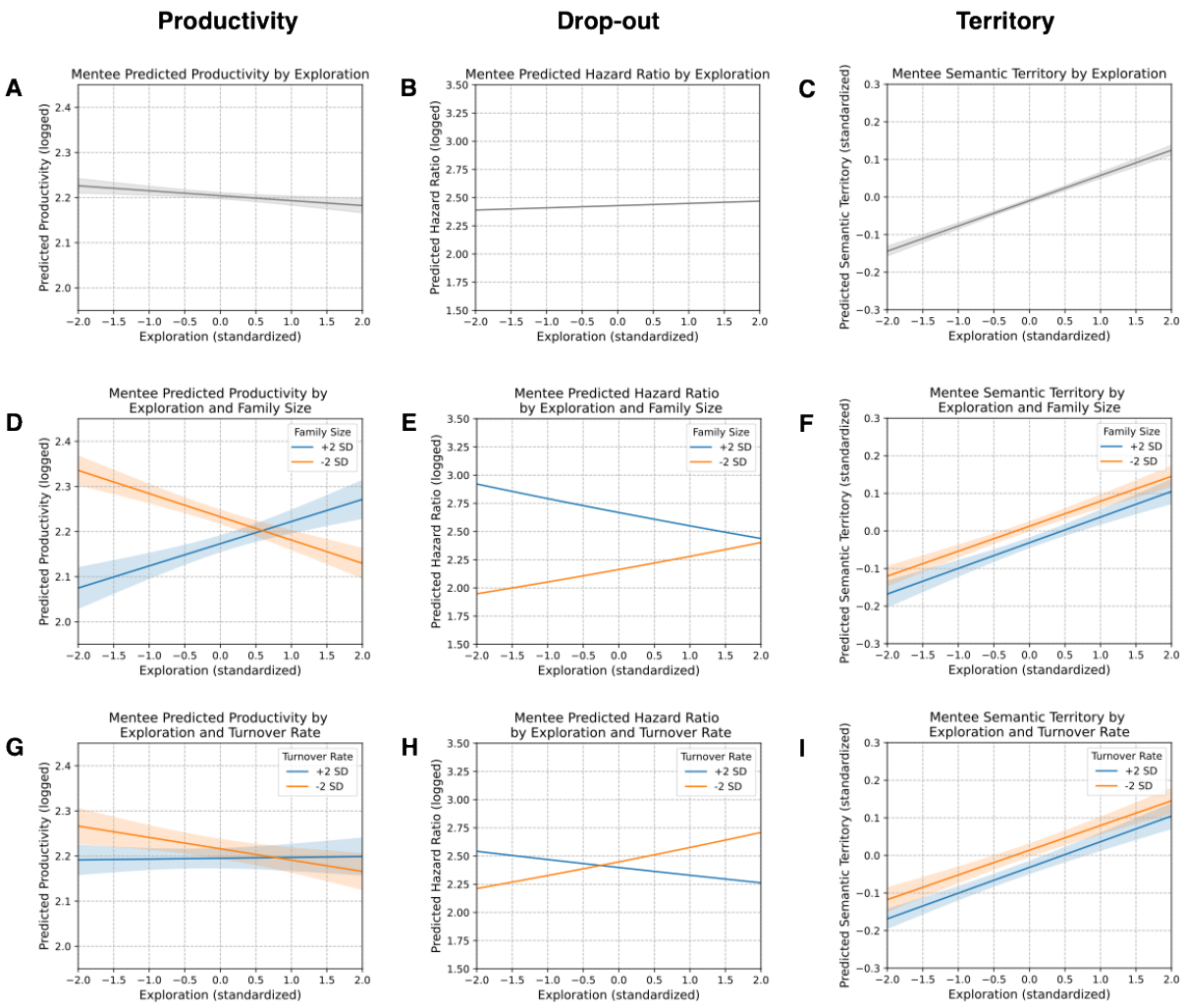

Fig.6. Exploration and Mentees Outcomes. Panels show model-predicted mentee outcomes as a function of exploration (standardized). (A,D,G) The main effects of exploration on mentee productivity (log publications), the hazard of exiting academia, and mentees' semantic territory, respectively. (B,E,H) Predicted outcomes by exploration interacted with family size (±2 SD). (C,F,I) Predicted outcomes by exploration interacted with field dynamics, measured as conceptual turnover (±2 SD). Shaded bands indicate 90% confidence intervals. All predictions are based on the full multivariate models reported in the Table S4-S6 in SI, with continuous covariates held at their means and indicator variables set to their reference categories. We further replicated the mentee-level regressions by substituting the family-level exploration measure with each mentee's own deviation from their mentor's prior research, and find that the results are consistent with the aforementioned findings (Table S7-9 and Fig.S8 in SI).

References

1. B. Bozeman, E. Corley, Scientists' collaboration strategies: implications for scientific and technical human capital. *Res. Policy* **33**, 599-616 (2004).
2. G. Laudel, J. Glaser, From apprentice to colleague: The metamorphosis of early career researchers. *Higher Ed.* **55**, 387-406 (2008).
3. P. E. Stephan, *How economics shapes science*. (Harvard University Press, Cambridge, MA, 2012).
4. R. D. Malmgren, J. M. Ottino, L. A. N. Amaral, The role of mentorship in protege performance. *Nature* **465**, 622-U117 (2010).
5. S. V. David, B. Y. Hayden, Neurotree: A Collaborative, Graphical Database of the Academic Genealogy of Neuroscience. *PLoS One* **7**, (2012).
6. W. Dores, F. Benevenuto, A. H. F. Laender, Ieee, *Extracting Academic Genealogy Trees from the Networked Digital Library of Theses and Dissertations*. 2016 Ieee/Acm Joint Conference on Digital Libraries (Ieee, New York, 2016), pp. 163-166.
7. S. Shibayama, Sustainable development of science and scientists: Academic training in life science labs. *Res. Policy* **48**, 676-692 (2019).
8. T. Jia, D. Wang, B. K. Szymanski, Quantifying patterns of research-interest evolution. *Nature Human Behaviour* **1**, 0078 (2017).
9. W. D. Hamilton, R. M. May, DISPERSAL IN STABLE HABITATS. *Nature* **269**, 578-581 (1977).
10. S. Delamont, P. Atkinson, O. Parry, Critical mass and doctoral research: reflections on the Harris report. *Studies in Higher Education* **22**, 319-331 (1997).
11. S. Shibayama, Y. Baba, J. P. Walsh, Organizational Design of University Laboratories: Task Allocation and Lab Performance in Japanese Bioscience Laboratories. *Res. Policy* **44**, 610–622 (2015).
12. L. Rossi, I. L. Freire, J. P. Mena-Chalco, Genealogical index: A metric to analyze advisor-advisee relationships. *J. Informetr.* **11**, 564-582 (2017).
13. R. J. P. Damaceno, L. Rossi, R. Mugnaini, J. P. Mena-Chalco, The Brazilian academic genealogy: evidence of advisor-advisee relationships through quantitative analysis. *Scientometrics* **119**, 303-333 (2019).
14. L. Rossi, R. J. P. Damaceno, I. L. Freire, E. J. H. Bechara, J. P. Mena-Chalco, Topological metrics in academic genealogy graphs. *J. Informetr.* **12**, 1042-1058 (2018).
15. J. Wang, S. Shibayama, Mentorship and creativity: Effects of mentor creativity and mentoring style. *Res. Policy* **51**, 104451 (2022).
16. H. W. Marsh, K. J. Rowe, A. Martin, PhD students' evaluations of research supervision - Issues, complexities, and challenges in a nationwide Australian experiment in benchmarking universities. *J. of Higher Ed.* **73**, 313-348 (2002).
17. G. BROWN, M. ATKINS, *Effective Teaching in Higher Education*. (Methuen, London, 1988).
18. J. Hockey, THE SOCIAL-SCIENCE PHD - A LITERATURE-REVIEW. *Studies in Higher Education* **16**, 319-332 (1991).
19. W. Bastalich, Content and context in knowledge production: a critical review of doctoral supervision literature. *Studies in Higher Education* **42**, 1145-1157 (2017).
20. J. Hockey, Motives and meaning amongst PhD supervisors in the social sciences. *British Journal of Sociology of Education* **17**, 489-506 (1996).
21. Z. Lin, Y. Yin, L. Liu, D. Wang, SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data* **10**, 315 (2023).

Supporting Information

S1. Dataset Source and Sample Construction

S1.1 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 supervisor names. PQDT Global provides structured metadata for 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), 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 enables us to trace complete academic lineages from their point of origin, preventing left-censoring in family-level 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 PhD records in our dataset with data from the NSF's Survey of Earned Doctorates (SED). As shown in Fig. S1.C, our dataset closely follows the SED data, with only minor discrepancies in a few years due to missing records.

S1.2 Bibliometric Data

To identify publication activities of academic families (both mentors and mentees), we linked publication records to all authors in our mentor-mentee pair dataset constructed from PQDT Global (see S.2). 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 (*1*), 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.

S.2 Construction of Mentor–Mentee Links (Data A)

S.2.1 Assigning Mentees to Mentors

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. S1.D). Among dissertation records that report supervisor information, approximately 90% list a single supervisor, while the remaining 10% list two or more supervisors (Fig. S2.A). 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. S2.B).

S.2.2 Disambiguating Student Authors and Supervisors.

While mentor–mentee relationships can be directly inferred by linking student names and supervisor names in PQDT Global dissertation records, reconstructing academic families from these records requires identifying individuals who transition from mentees to mentors over time. Specifically, we identify formal doctoral students who later appear as supervisors on subsequent dissertations. Linking these records allows us to reconstruct academic family lineages and to trace the number of students trained by former mentees once they themselves become supervisors. We address this by identifying individuals who appear both as student authors and, in later years, as supervisor authors in PQDT Global, and we implement a multi-step matching strategy to link student and supervisor records belonging to the same individual.

The first necessary condition for matching is that a student author and a supervisor author share the same name (Fig. S3.A). 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 (*2*). 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 pairing. 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 (*3*). 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, corresponding to the career stage in which faculty are most likely to actively supervise doctoral students—following completion of doctoral and postdoctoral training and spanning the pre- and post-tenure periods in U.S. academia (*4*). Intervals between 5 and 8 years received 1 point, reflecting early-career supervision. 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 (*5*) 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 S14.

S.2.3 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 match scores greater than or equal to 4 (Fig. S4.A). 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 a few highly productive mentors, around the top 5% in the distribution, who have produced more than ten mentees (Fig.S4.B).

S.3 Linking PQDT Global and SciSciNet (Data B)

We linked publication records in SciSciNet to individuals identified in mentor–mentee pairs constructed from PQDT Global dissertation records. Because all mentors in our dataset also appear as former dissertation authors, our objective was to match each dissertation author—both mentees and mentors—to their corresponding publication records in SciSciNet (Fig. S3B). This linkage enables us to connect dissertation-based academic genealogies to individuals' downstream publication trajectories.

S.3.1 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.

S.3.2 Scoring and Matching

For each candidate dissertation–publication author pair, we computed a composite match score by summing points across multiple criteria (Table S15), 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 (*5*), 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.

S.3.3 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 resulted in 720,657 unique PQDT Global authors. For comparison, we began with 1,059,244 US PhD dissertation records from the PQDT Global dataset. This suggests that, at a minimum, approximately 68% of PhD graduates have their doctoral dissertations published (Fig.S5.A).

S.4 Mering mentor–mentee pairs (Dataset A) with publication records (Dataset B)

S.4.1. Merged dataset

We combine two datasets: mentor–mentee pairs identified from the PQDT Global dissertation records (Dataset A) and matched PQDT Global author–SciSciNet author pairs (Dataset B). Merging these datasets allows us to observe both mentor–mentee relationships and the subsequent publication trajectories. 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.

S.4.2. Final analysis sample (used in main regressions)

The merged dataset is further restricted to observations with non-missing values for all variables required in the main regression models, including exploration, family size, field dynamics, control variables, and outcome measures. After this restriction, the final analysis sample consists of 17,010 academic families (mentors) at the mentor level and a corresponding 78,102 mentees at the mentee level. This sample forms the basis of all regression analyses reported in the main manuscript and SI Tables S1–S6. For the mentee-level analysis, we restrict the sample to mentees from academic families with family size ≤ 4. After applying this restriction and excluding observations with missing covariates, the final mentee-level analysis sample consists of 51,363 mentees.

S.5 Measuring Exploration

S.5.1 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 (*6*). 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:

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

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

S.5.2 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 family's exploration strategy. For each family *f*, we calculate the average of the mentee's median distances:

- $Exploration_f = \frac{1}{n_f}\sum_{i\in\mathcal{S}_f} d_i$

where Exploration_f captures the exploration of the family *f*, $\mathcal{S}_f$ is the set of mentees from academic family *f*. Lastly, the family size n_f is the number of mentees trained by a given family, which serves as a proxy for the size of the mentor's academic family.

Of the 168,647 mentor-mentee pairs (corresponding to 168,647 unique mentees), we computed a semantic distance measure for 134,638 mentees (Fig. S6.A). 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 family-level exploration measure by averaging across all mentees associated with each mentor. This aggregation yielded 39,117 unique mentors with exploration measures (Fig.S6.B).

S.6 Field Dynamics

S.6.1 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 (*7, 8*). 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.

S.6.2 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.

S.7 Measurement of Semantic Territory

S.7.1 Family-level 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 (*6*). 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.S7.A). 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.

S.7.2 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.S7.B) 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

S.8 Econometric Specifications

S.8.1 Family-level Specification

To examine how exploration relates to family-level outcomes, we estimate the following baseline specification:

$$Y_j = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + Z_j'\delta + \gamma_{c(j)} + \lambda_{f(j)} + \varepsilon_j$$

We additionally estimate models including interaction terms:

$$\beta_4\left(\text{Exploration}_j \times \text{FamilySize}_j\right) \text{ and } \beta_4\left(\text{Exploration}_j \times \text{FieldDynamics}_j\right)$$

Y_j denotes our primary outcome of interest for family *j*, which includes (1) the mentor's own productivity (total number of publications at the career-level), (2) the sustainability of the family lineage (number of mentees produced by a family who later become mentors themselves), and (3) the expansion of the semantic territory (coverage of semantic space by the family). Our key explanatory variable, exploration, $\left(exploration_j\right)$ 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_j', including mentors' average citation impact and whether they have graduated from elite universities (Ivy Plus Universities[2]). We also include cohort fixed effects, $\gamma_{c(j)}$, based on the mentor's PhD dissertation year, accounting for potential temporal differences across generations, as well as field fixed effects, $\lambda_{f(j)}$, to control for discipline-specific baseline differences. Fields were derived from the most common MAG field categories in the mentor's publications.

S.8.2 Mentee-level Specification

For the mentee-level analysis, we estimate family-level random effect models to account for the nested structure of the data, in which mentees are clustered within families. We denote Y_{ij} as the outcome for mentee *i*, trained under family *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). We estimate:

$$Y_{ij} = \alpha + \beta_1 \text{Exploration}_j + \beta_2 \text{FamilySize}_j + \beta_3 \text{FieldDynamics}_j + Z_{ij}'\delta + u_j + \varepsilon_{ij}$$

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

where $u_j \sim \mathcal{N}(0, \sigma_u^2)$ represents a family-level random intercept capturing unobserved heterogeneity shared by mentees within the same academic family. As in the family-level analysis, we further estimate models including interaction terms between exploration and family size, and between exploration and field dynamics.

As control variables, the vector of $Z_{i,j}$, includes indicators for whether the mentor obtained a PhD from an elite university, whether the mentee obtained a PhD from an elite university, and the mentor's average citation impact. Similar to family-level analysis, we also include cohort fixed effects, $\gamma_{c(i)}$, based on the mentee's PhD dissertation year, and field fixed effects, $\lambda_{f(i)}$, to control for discipline-specific baseline differences.

Figures

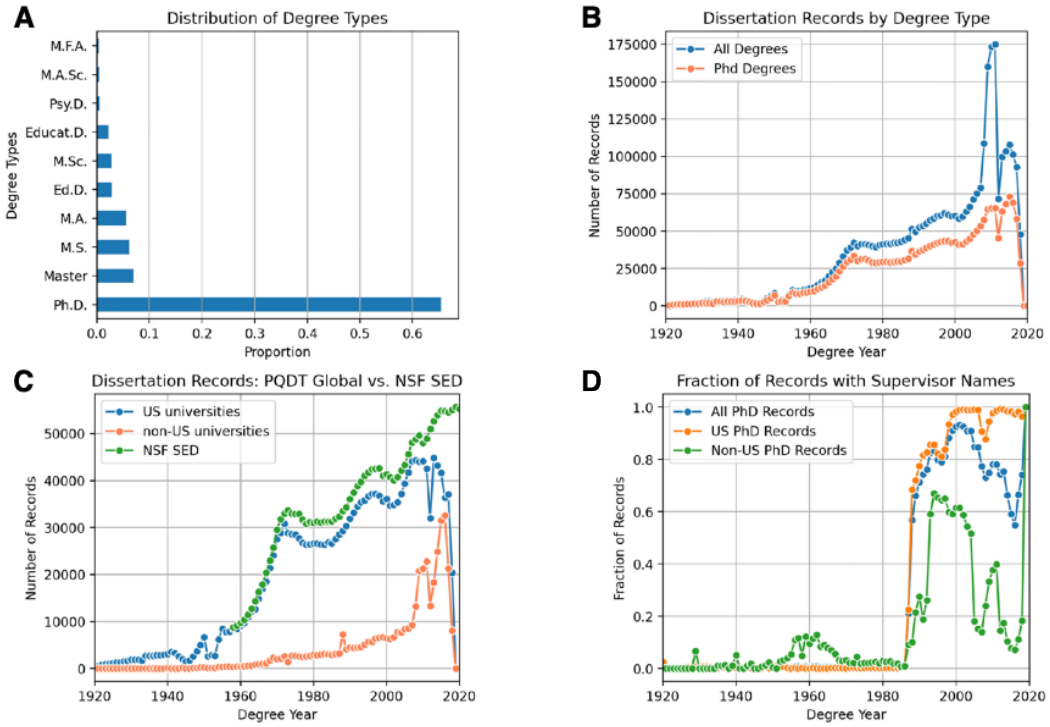

Fig.S1. Coverage and metadata availability in ProQuest Dissertations & Theses Global (PQDT Global). Panels A–D summarize the composition, temporal coverage, and supervisor coverage of dissertation records in PQDT Global. (A) Distribution of degree types. (B) Annual counts of dissertations by degree. (C) Temporal trends in dissertation records with and without supervisor name information. (D) Fraction of dissertations with identifiable supervisors by year, illustrating the rapid increase in supervisor coverage in recent cohorts.

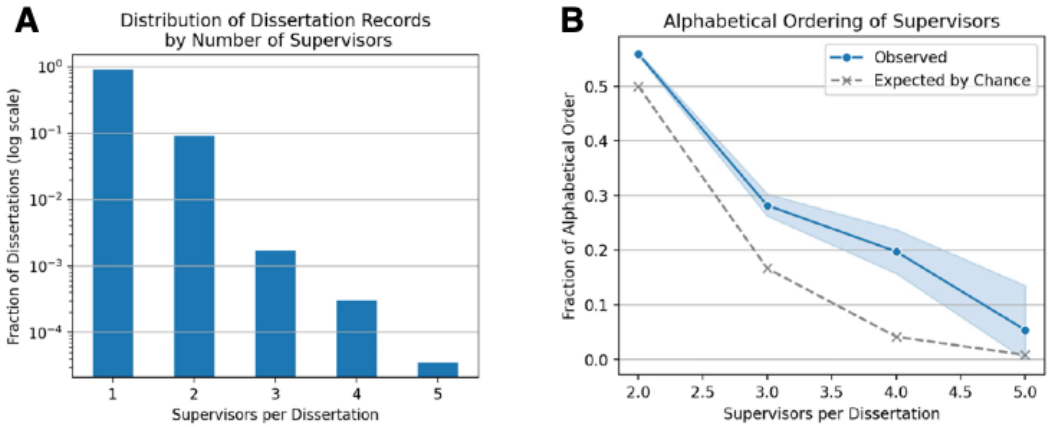

Fig.S2. Supervisor counts and ordering in dissertation records. (A) Distribution of dissertation records by the number of supervisors per dissertation; more than 90% of records list a single supervisor. (B) Fraction of dissertation records in which supervisors are listed in alphabetical order, by the 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.

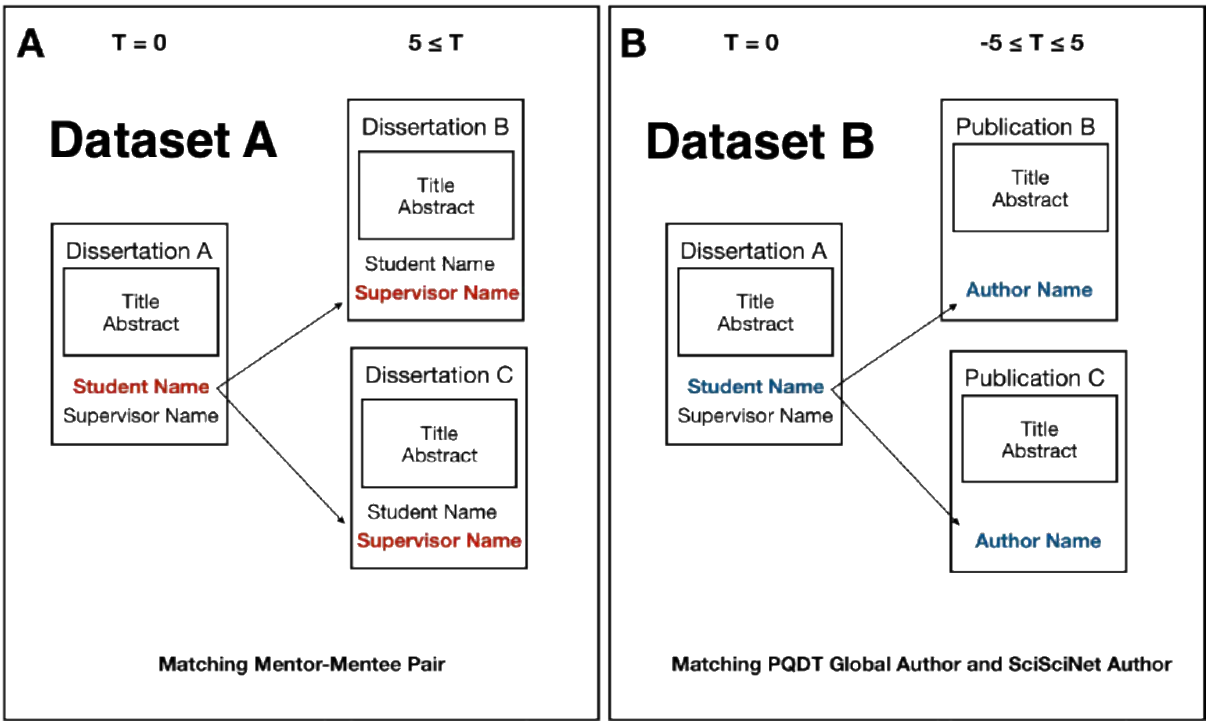

Fig.S3. Linking mentor–mentee relationships and publications across data sources. (A) Construction of the Dataset A, mentor–mentee pairs within ProQuest Dissertations & Theses Global (PQDT Global) by matching student names in dissertation records to supervisor names and identifying individuals who later appear as supervisors. (B) Construction of Dataset B, by linking PQDT Global authors to publication records by matching dissertation authors to author identities in SciSciNet.

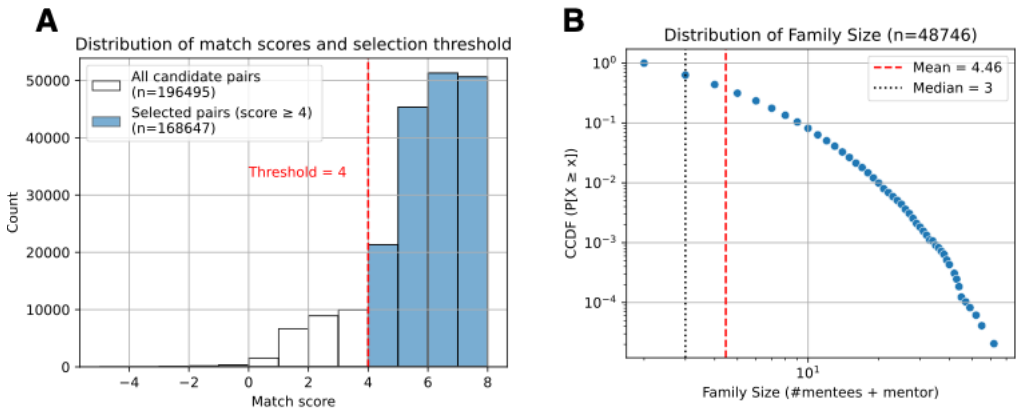

Fig. S4. Distribution of mentor–mentee matching scores and the family size. (A) Distribution of mentor–mentee matching scores based on the rules listed in Table S1, with the vertical line indicating the threshold used for pair selection. (B) Distribution of the number of mentees per family (including mentor) after applying the matching procedure.

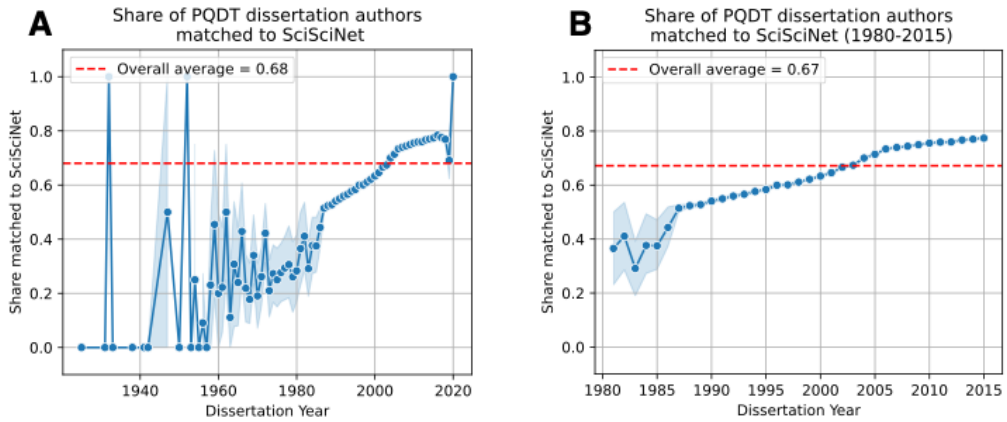

Fig.S5. Share of PQDT Global dissertation authors matched to SciSciNet authors. (A) Annual share of PQDT Global Dissertation authors that are matched to SciSciNet authors. (B) Same measure restricted to the period 1980-2015.

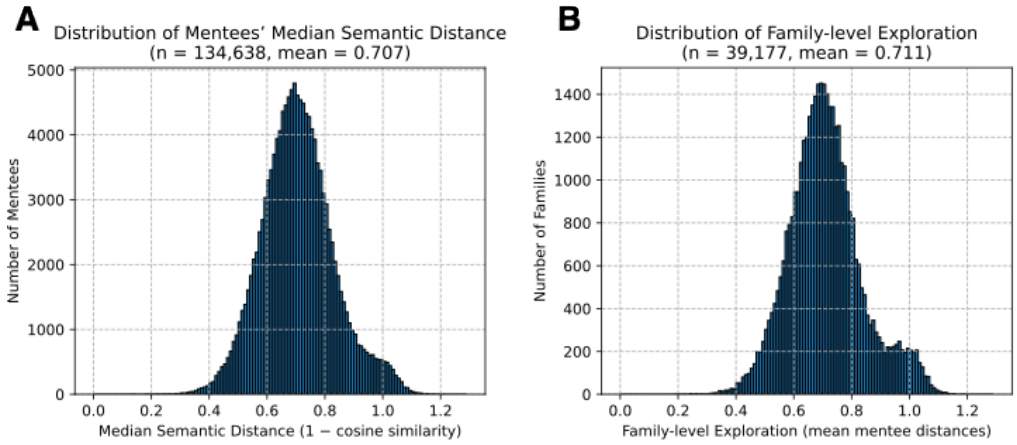

Fig.S.6. Distributions of mentee semantic distance and mentor exploration. (A) Distribution of mentee's median semantic distance. (B) Distribution of family-level exploration.

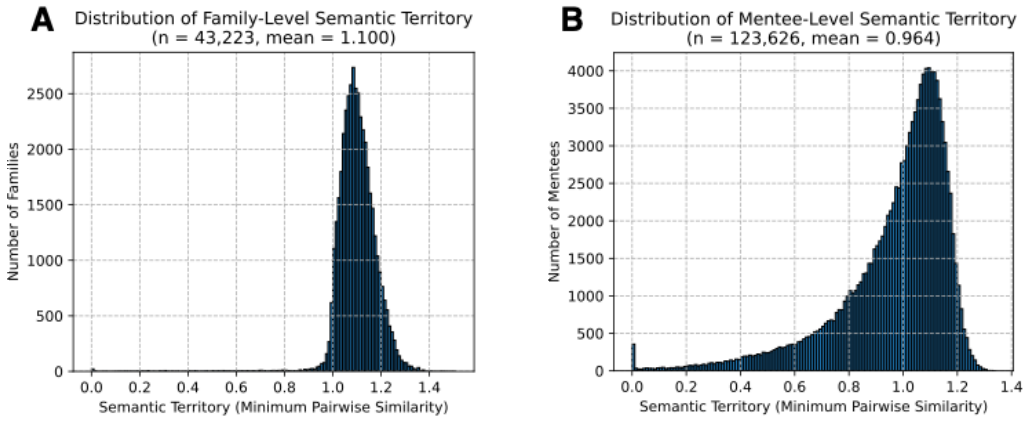

Fig.S.7. Distributions of family-level and mentee-level semantic territory. (A) Distribution of family-level semantic territory. (B) Distribution of mentee-level semantic territory.

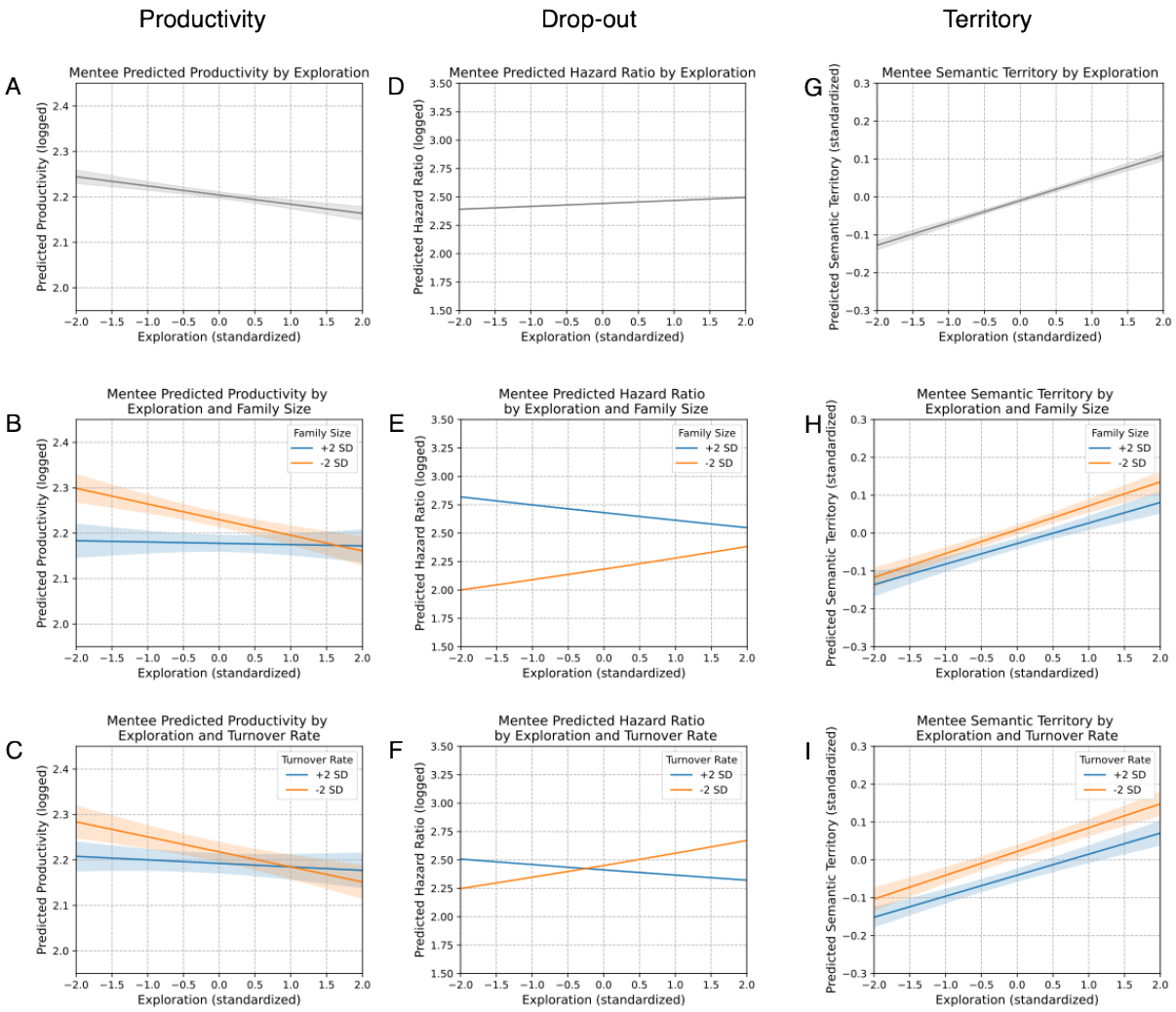

Fig.S.8. Mentee-level Exploration and Mentees' Outcomes. Panels show model-predicted mentee outcomes as a function of exploration (standardized). (A,D,G) The main effects of the mentee's own exploration on mentee productivity (log publications), the hazard of exiting academia, and mentees' semantic territory, respectively. (B,E,H) Predicted outcomes by exploration interacted with family size (±2 SD). (C,F,I) Predicted outcomes by exploration interacted with field dynamics, measured as conceptual turnover (±2 SD). Shaded bands indicate 90% confidence intervals. All predictions are based on the full multivariate models reported in the Table S7-S9, with continuous covariates held at their means and indicator variables set to their reference categories.

Tables

Table S1. OLS Regressions Predicting Mentor Productivity

	Dependent variable:		
	Number of Papers		
	(1)	(2)	(3)
exploration(std)	-0.236***	-0.226***	-0.237***
	(0.014)	(0.051)	(0.014)
family size(log)	0.596***	0.597***	0.596***
	(0.023)	(0.022)	(0.023)
field dyanmics(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 c10(log)	0.162***	0.162***	0.162***
	(0.017)	(0.017)	(0.017)
exploration(std)×family size(log)		-0.006	
		(0.031)	
exploration(std)×field dynamics(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
F Statistic	66.637***	65.401***	65.401***

Note: Estimates are from ordinary least squares (OLS) regressions. Robust standard errors are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S2. Poisson Regressions Predicting Number of Future Mentors

	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(log)	1.403***	1.401***	1.402***
	(0.033)	(0.033)	(0.033)
field dyanmics(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 c10(log)	0.235***	0.235***	0.234***
	(0.026)	(0.026)	(0.026)
exploration(std)×family size(log)		0.016	
		(0.038)	
exploration(std)×field dynamics(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: Estimates are from Poisson regression models. Robust standard errors are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S3. OLS Regressions Predicting Mentor Semantic Territory

	Dependent variable:		
	Semantic Territory		
	(1)	(2)	(3)
exploration(std)	0.044***	-0.106***	0.041***
	(0.008)	(0.029)	(0.008)
family size(log)	0.761***	0.755***	0.761***
	(0.016)	(0.016)	(0.016)
field dyanmics(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 c10(log)	0.078***	0.080***	0.078***
	(0.009)	(0.009)	(0.009)
exploration(std)×family size(log)		0.091***	
		(0.017)	
exploration(std)×field dynamics(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
F Statistic	92.971***	92.005***	91.412***

Note: Estimates are from ordinary least squares (OLS) regressions. Robust standard errors are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S4. OLS Regressions Predicting Mentee Productivity

	Dependent variable:		
	Number of Papers		
	(1)	(2)	(3)
exploration(std)	-0.011**	-0.101***	-0.012**
	(0.005)	(0.020)	(0.005)
family size(log)	-0.025***	-0.028***	-0.025***
	(0.010)	(0.010)	(0.010)
field dyanmics(std)	-0.007	-0.006	-0.005
	(0.007)	(0.006)	(0.007)
career_length	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 c10(log)	0.183***	0.183***	0.183***
	(0.007)	(0.007)	(0.007)
exploration(std)×family size(log)		0.047***	
		(0.010)	
exploration(std)×field dynamics(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
F Statistic	96,659.800***	96,740.800***	96,665.030***

Note: Estimates are from ordinary least squares (OLS) random-effects models with random intercepts at the academic family level. Robust standard errors clustered at the family level are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S5. Cox Regressions Predicting Mentee's Exit Hazards

	Dependent variable:		
	Exit Hazards		
	(1)	(2)	(3)
exploration(std)	0.008	0.101***	0.011*
	(0.006)	(0.022)	(0.006)
family size(log)	0.095***	0.098***	0.095***
	(0.010)	(0.011)	(0.010)
field dyanmics(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 c10(log)	-0.187***	-0.188***	-0.188***
	(0.007)	(0.007)	(0.007)
exploration(std)×family size(log)		-0.046***	
		(0.010)	
exploration(std)×field dynamics(std)			-0.020***
			(0.005)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	68,336	68,336	68,336
Log Likelihood	-413,834.000	-413,824.300	-413,827.400
LR Test	2,364.337***	2,383.793***	2,377.581***

Note: Estimates are from Cox proportional hazards regression models. Robust standard errors are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S6. OLS Regressions Predicting Mentee Semantic Territory

	Dependent variable:		
	Semantic Territory		
	(1)	(2)	(3)
exploration(std)	0.067***	0.065***	0.067***
	(0.004)	(0.016)	(0.004)
family size(log)	-0.021***	-0.021***	-0.021***
	(0.007)	(0.007)	(0.007)
field dyanmics(std)	-0.012**	-0.012**	-0.011**
	(0.005)	(0.005)	(0.005)
total publications(log)	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 c10(log)	0.008	0.008	0.008
	(0.005)	(0.005)	(0.005)
exploration(std)×family size(log)		0.001	
		(0.008)	
exploration(std)×field dynamics(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
F Statistic	64,222.750***	64,221.120***	64,221.200***

Note: Estimates are from ordinary least squares (OLS) random-effects models with random intercepts at the academic family level. Robust standard errors clustered at the family level are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S7. OLS Regressions Predicting Mentee Productivity

	Dependent variable:		
	Number of Papers		
	(1)	(2)	(3)
mentee exploration(std)	-0.020***	-0.050***	-0.020***
	(0.004)	(0.017)	(0.004)
family size(log)	-0.024**	-0.024**	-0.024**
	(0.010)	(0.010)	(0.010)
field dyanmics(std)	-0.008	-0.008	-0.006
	(0.006)	(0.006)	(0.007)
career length	0.143***	0.143***	0.143***
	(0.001)	(0.001)	(0.001)
mentee elite univ	0.062***	0.062***	0.061***
	(0.014)	(0.014)	(0.014)
mentor elite univ	-0.001	-0.001	-0.001
	(0.010)	(0.010)	(0.010)
mentor average c10(log)	0.181***	0.181***	0.181***
	(0.007)	(0.007)	(0.007)
exploration(std)×family size(log)		0.015*	
		(0.008)	
exploration(std)×field dynamics(std)			0.006
			(0.004)
Constant	-1.476***	-1.467***	-1.476***
	(0.102)	(0.103)	(0.102)
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
F Statistic	96,707.540***	96,732.740***	96,713.720***

Note: Estimates are from ordinary least squares (OLS) random-effects models with random intercepts at the academic family level. Robust standard errors clustered at the family level are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S8. Cox Regressions Predicting Mentee's Exit Hazards

	Dependent variable:		
	Exit Hazards		
	(1)	(2)	(3)
exploration(std)	0.011**	0.078***	0.012**
	(0.005)	(0.022)	(0.005)
family size(log)	0.094***	0.096***	0.095***
	(0.010)	(0.010)	(0.010)
field dyanmics(std)	-0.001	-0.002	-0.004
	(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***	-0.058***
	(0.013)	(0.013)	(0.013)
mentor average c10(log)	-0.187***	-0.187***	-0.188***
	(0.007)	(0.007)	(0.007)
exploration(std)×family size(log)		-0.032***	
		(0.010)	
exploration(std)×field dynamics(std)			-0.016***
			(0.005)
Cohort FE	Yes	Yes	Yes
Fields FE	Yes	Yes	Yes
Observations	68,336	68,336	68,336
Log Likelihood	-413,833.100	-413,827.900	-413,828.900
LR Test	2,366.179***	2,376.517***	2,374.570***

Note: Estimates are from Cox proportional hazards regression models. Robust standard errors are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S9. OLS Regressions Predicting Mentee Semantic Territory

	Dependent variable:		
	Semantic Territory		
	(1)	(2)	(3)
exploration(std)	0.059***	0.067***	0.059***
	(0.004)	(0.015)	(0.004)
family size(log)	-0.017**	-0.017**	-0.017**
	(0.007)	(0.007)	(0.007)
field dynamics(std)	-0.015***	-0.015***	-0.016***
	(0.005)	(0.005)	(0.005)
total publications(log)	0.645***	0.645***	0.645***
	(0.004)	(0.004)	(0.004)
mentee elite univ	0.001	0.001	0.001
	(0.010)	(0.010)	(0.010)
mentor elite univ	0.006	0.006	0.006
	(0.008)	(0.008)	(0.008)
mentor average c10(log)	0.003	0.003	0.003
	(0.005)	(0.005)	(0.005)
exploration(std)×family size(log)		-0.004	
		(0.007)	
exploration(std)×field dynamics(std)			-0.002
			(0.004)
constant	-1.652***	-1.655***	-1.653***
	(0.135)	(0.136)	(0.135)
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
F Statistic	64,134.090***	64,133.330***	64,132.970***

Note: Estimates are from ordinary least squares (OLS) random-effects models with random intercepts at the academic family level. Robust standard errors clustered at the family level are reported in parentheses. All models include cohort and field fixed effects. *p<0.1; **p<0.05; ***p<0.01

Table S10. Mentor-level Summary Statistics

Variable Names	Descriptions	N	Mean	SD	Min	Median	Max
(1) number of papers	Total career publications of the mentor, measured as the cumulative number of papers indexed in SciSciNet.	17010	111.44	136.12	1	78	2573
(2) number of future supervisors	Number of a mentor's mentees who later become mentors themselves.	17010	0.21	0.58	0	0	8
(3) semantic territory	Breadth of the academic family's research topics, measured as the maximum semantic distance among family-level publications.	17010	1.14	0.06	0.94	1.13	1.45
(4) exploration	Family-level exploration, measured as the mean of mentee-level semantic distances between dissertations and the mentor's prior publications.	17010	0.70	0.11	0.27	0.70	1.14
(5) family size	Total number of individuals in the academic family, defined as the number of mentees supervised by the mentor plus the mentor.	17010	7.41	3.84	4	6	63
(6) field dynamics	Conceptual turnover in the mentor's field, measured as the proportion of nonoverlapping noun phrases between consecutive time periods.	17010	0.62	0.03	0.52	0.61	0.75
(7) mentor elite univ	Indicator variable equal to 1 if the mentor received their PhD from an elite university, and 0 otherwise.	17010	0.20	0.40	0	0	1
(8) mentor average c10 log	Log-transformed average number of citations within 10 years (C10) across the mentor's publications prior to training the focal cohort.	17010	2.80	0.91	0.00	2.85	7.93

Table S11. Mentor-level Correlation Table

Variable Names	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) number of papers							
(2) number of future supervisors	0.170						
(3) semantic territory	0.460	0.278					
(4) exploration	-0.039	0.006	0.050				
(5) family size	0.293	0.428	0.354	0.008			
(6) field dynamics	0.136	-0.025	0.096	-0.153	0.071		
(7) mentor elite univ	0.045	0.025	0.122	-0.030	0.032	0.076	
(8) mentor average c10 log	0.055	0.048	0.019	-0.287	0.062	0.012	0.078

Table S12. Mentee-level Summary Statistics

Variable Names	Descriptions	N	Mean	SD	Min	Median	Max
(1) number of papers	Total career publications of the mentee, measured as the cumulative number of papers indexed in SciSciNet.	68336	24.60	61.54	1	9	4381
(2) career length	Number of years between the mentee's first publication and the last observed publication year, right-censored at 2018.	68336	8.99	7.28	1	8	61
(3) career exit	Event indicator equal to 1 if the mentee's last observed publication occurs before 2018 (exit event), and 0 if the publication record is right-censored at 2018.	68336	0.57	0.49	0	1	1
(4) semantic territory	Breadth of the mentee's research topics, measured as the maximum semantic distance among the mentee's publications.	51363	0.94	0.21	0.00	0.99	1.33
(5) exploration	Exploration measure of the academic family, defined as the mean semantic distance between mentees' dissertations and the mentor's prior publications.	68336	0.70	0.09	0.27	0.70	1.10
(6) exploration mentee	Individual deviation measure, defined as the semantic distance between the mentee's dissertation and the mentor's prior publications.	68336	0.70	0.12	0.19	0.69	1.16
(7) family size	Total number of individuals in the academic family (number of mentees supervised by the mentor plus the mentor).	68336	9.78	6.38	4	8	62
(8) field dynamics	Conceptual turnover in the mentor's field, measured as the proportion of nonoverlapping noun phrases between consecutive time periods.	68336	0.62	0.03	0.53	0.62	0.75
(9) mentee elite univ (mentee)	Indicator equal to 1 if the mentee received their PhD from an elite university, and 0 otherwise.	68336	0.11	0.31	0	0	1
(10) mentor elite univ	Indicator equal to 1 if the mentor received their PhD from an elite university, and 0 otherwise.	68336	0.21	0.41	0	0	1
(11) mentor average c10 log	Log-transformed average number of citations within 10 years (C10) across the mentor's publications prior to supervising the mentee.	68336	2.87	0.81	0.00	2.89	7.00

Table S13. Mentee-level Correlation Table

Variable Names	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) number of papers (mentee										
(2) career length (mentee)	0.456									
(3) career exit (mentee)	-0.342	-0.578								
(4) semantic territory (mentee)	0.340	0.556	-0.564							
(5) exploration (family)	0.002	-0.018	0.025	0.076						
(6) exploration (mentee)	0.001	-0.022	0.017	0.069	0.815					
(7) family size (family)	0.030	0.025	0.054	-0.004	0.058	0.046				
(8) field dynamics (family)	0.025	0.024	0.041	-0.051	-0.183	-0.144	0.120			
(9) mentee elite univ (mentee)	0.045	0.064	-0.048	0.046	-0.001	-0.002	0.075	0.079		
(10) mentor elite univ (family)	0.018	0.035	-0.033	0.028	-0.024	-0.020	0.031	0.073	0.258	
(11) mentor average c10 log (family)	0.012	0.056	-0.067	0.000	-0.321	-0.262	0.066	0.055	0.147	0.106

Table S14. 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 S15. 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

References

1. Z. Lin, Y. Yin, L. Liu, D. Wang, SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data* **10**, 315 (2023).
2. S. Milojević, Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics* **7**, 767-773 (2013).
3. S. National Center for, S. Engineering. (National Science Foundation, 2024).
4. G. Tripodi *et al.*, Tenure and research trajectories. *Proceedings of the National Academy of Sciences* **122**, e2500322122 (2025).
5. A. Singh, M. D'Arcy, A. Cohan, D. Downey, S. Feldman, Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*, (2022).
6. N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, (2019).
7. S. Milojević, Quantifying the cognitive extent of science. *Journal of Informetrics* **9**, 962-973 (2015).
8. M. Cheng *et al.*, How New Ideas Diffuse in Science. *American Sociological Review* **88**, 522-561 (2023).