Research on Knowledge Inheritance Within Academic Lineages Based on the LDA Topic Model: A Case Study of the Genetics Academic Lineage Centered on Tan Jiazhen (Postprint)

Authors: Liu Junwan, Yang Bo, Wang Feifei, Xu Shuo

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] Academic genealogies propel scientific development through knowledge inheritance. Investigating the characteristics of knowledge inheritance and exploring the inheritance patterns of academic genealogies and their impact effectiveness on academic output provides a reference for understanding talent growth patterns and formulating talent policies. [Method/Process] Based on the LDA topic model, this study examines journal articles published by members of the academic genealogy centered on Tan Jiazhen in the field of genetics, extracts research topics of these genealogy members, adopts the biological concepts of 'inheritance' and 'mutation', classifies genealogy members into 'inheritance scholars', 'mutation scholars', and 'non-inheritance non-mutation scholars' according to topic similarity, and analyzes the academic performance of these three types of scholars. [Results/Conclusions] The analysis results indicate that 'inheritance scholars' and 'mutation scholars' within Tan Jiazhen's academic genealogy demonstrate relatively high academic performance; 'non-inheritance non-mutation scholars' constitute the largest proportion numerically, yet exhibit relatively low academic performance; the distribution of 'inheritance scholars' and 'mutation scholars' across different topics shows significant differences.

Full Text

Preamble

Vol. 62 No. 10, May 2018, ChinaXiv Cooperative Journal

Research on Knowledge Inheritance within Academic Pedigrees Based on LDA Topic Models

— A Case Study of the Genetics Academic Pedigree with Tan Jiazhen as the Core

Liu Junwan, Yang Bo, Wang Feifei, Xu Shuo
School of Economics and Management, Beijing University of Technology, Beijing 100124

Abstract

[**Purpose/Significance**] Academic pedigrees promote scientific development through knowledge inheritance. Studying the characteristics of knowledge transmission, exploring inheritance patterns within academic pedigrees and their impact on academic output provides valuable references for understanding talent development patterns and formulating talent policies.

[**Method/Process**] Using the LDA topic model, this study examines journal articles published by members of Tan Jiazhen's genetics academic pedigree. Research topics of pedigree members are extracted, and drawing upon the biological concepts of "heredity" and "variation," scholars are classified into "hereditary scholars," "variation scholars," and "non-hereditary non-variation scholars" based on topic similarity. The academic performance of these three types of scholars is then analyzed.

[**Results/Conclusions**] The results indicate that "hereditary scholars" and "variation scholars" within Tan Jiazhen's academic pedigree demonstrate relatively higher academic performance. While "non-hereditary non-variation scholars" constitute the largest proportion, their academic performance is relatively low. The distribution of "hereditary scholars" and "variation scholars" across different topics shows significant differences.

Keywords: academic pedigree, knowledge inheritance, topic model, genetics

1 Introduction

"Life is finite, but knowledge is infinite." In the face of the world's infinite mysteries, individual power is minuscule. Yet, as streams converge to form vast oceans, knowledge accumulates through human generations and transmits across pedigrees and schools, forging today's flourishing scientific system. The history of Chinese science and technology, when examined at the level of individual research fields, represents the establishment, expansion, and evolution of academic pedigrees founded by leading scholars in each discipline. An academic pedigree represents the temporal continuation and transmission of paradigms voluntarily recognized by the academic community, constituting a necessary condition for academic accumulation. Previous studies on academic pedigrees have typically focused on tracing the growth trajectories and academic origins of successful scholars. These qualitative studies articulate scholars' personal perspectives on information accumulation and utilization, but without sufficient

data support, they have not yielded universal laws of talent development. Although different research fields and academic pedigrees within the same field exhibit varying inheritance patterns, the top-down flow of knowledge remains constant. Is there a correlation between these differences and knowledge flow? With the emergence of evaluation metrics such as the "h-index" and academic genealogy databases like AcademicTree, quantitative analysis of academic pedigrees has become feasible. Consequently, some scholars have begun using research literature as an entry point for quantitative studies on academic pedigree evaluation. Since quantitative research on academic pedigrees emerged relatively recently and remains underdeveloped, current studies primarily focus on establishing scholar relationship networks through co-authorship and citation relationships in academic literature, conducting research at the system, group, and individual levels. For example, R.D. Malmgren et al. found that scholars exhibit stronger academic fecundity during the first two-thirds of their careers compared to the final third by measuring the number of students and their academic impact. Domestic scholars using social network analysis to examine the Quaternary academic pedigree's collaboration network characteristics and academic inheritance revealed that students remaining in the academic "home base" demonstrate more stable long-term collaboration with their mentors, with significantly higher collaboration intensity than students who leave for other institutions. C. Sugimoto et al. analyzed research field transition characteristics of pedigree members for interdisciplinary theory research. While these studies extensively examine external features such as network structures, they have not quantified the inheritance of ideas and knowledge represented by text within pedigrees. Therefore, this study selects Tan Jiazhen's genetics academic pedigree—one of the longest-developed, most comprehensively documented, and largest in modern Chinese genetics—as its research object. Using topic modeling methods, we investigate changes in research topics between mentors and students and before/after students' graduation, exploring the degree of topic variation among pedigree members and its relationship with research performance. This approach aims to reveal regular patterns of knowledge inheritance within pedigrees and identify talent development pathways, providing references for research management departments in formulating science and technology talent policies.

2 Research Methods and Data Acquisition

2.1 Research Methods and Technical Route for Knowledge Inheritance in Academic Pedigrees

This study constructs an academic pedigree tree using mentor-mentee relationship data, extracts feature words from scholars' paper titles, abstracts, and keywords, and builds a scholar feature word library based on the correspondence between papers and scholars. The optimal number of topics for the LDA (Latent Dirichlet Allocation) topic model is determined through ten-fold cross-validation. The LDA model then generates a fixed-dimensional topic distribu-

tion vector for each scholar, and JS (Jensen-Shannon) distance measures the distance between vectors of different scholars to obtain topic similarity. Scholars are subsequently classified into "hereditary scholars," "non-hereditary non-variation scholars," and "variation scholars." Finally, we analyze the academic performance and overall topic distribution of these different scholar types. The research framework, shown in Figure 1 [Figure 1: see original paper], comprises four components: data acquisition, data preprocessing, topic extraction and similarity calculation, and feature analysis.

Data acquisition involves two parts: pedigree data and literature data. Academic pedigree research requires extensive collection of historical materials to identify intergenerational relationships and construct the pedigree tree, with data authenticity being critical. Journal literature serves as the primary carrier of scholars' stage research achievements and the main form for updating, disseminating, and exchanging scientific knowledge within the academic community, representing important academic output. To study knowledge inheritance within the pedigree, we examine all Chinese journal articles and graduate theses published by pedigree members.

Data preprocessing primarily involves natural language processing tasks, as LDA topic models require document collections represented by feature words as input. This study uses two types of input samples: (1) each document in the collection corresponds to all papers of one scholar, used to analyze research direction differences between scholars and their mentors; (2) each document corresponds to all papers published by a scholar before or after graduation, used to analyze research direction differences before and after graduation. Documents in both samples have no sequential order differences, with research direction variations manifesting as differences in feature word sets at the text level. This preprocessing mainly includes text segmentation and stop word removal. During text segmentation, reasonable trade-offs need to be made between over-segmentation of feature words and segmentation granularity being too small. During stop word removal, interference from noise words such as modal particles, adverbs, prepositions, and conjunctions is eliminated.

The LDA topic model algorithm generates a research topic distribution vector for each scholar. Vector similarity can quantitatively describe changes in research directions between mentors and students and before/after graduation, as well as the degree of knowledge inheritance. As an unsupervised machine learning method, LDA identifies latent topic information in large document collections or corpora. To achieve our research objectives, we replace the document layer in the LDA model with an author layer represented by the paper collection of pedigree scholars, where each author represents the collection of topics, keywords, and abstracts from their papers. This transformation aligns with the principles of the AT (Author-Topic) Model. The model assumes that authors are mixed distributions over topics (Author-Topic), while topics are probability distributions over words (Topic-Word). This assumption projects author datasets onto a topic space, reducing time complexity in large-scale data processing. The

hyperparameters α and β, author-topic distribution , and topic-word distribution ϕ are all latent variables. For an author set D = {d_1, d_2, d_3, …, d_M} containing M authors distributed across K topics Z = {z_1, z_2, z_3, …, z_K}, where all feature words in the author text collection form a vocabulary W = {w_1, w_2, w_3, …, w_N}, the probability density function for each author is given by Formula (1):

$$P(w|\alpha,\beta) = \int P(\theta|\alpha)\left(\prod\sum P(z_n|\theta)P(w_n|z_n),\beta\right)d\theta \quad (1)$$

The above method aligns with the AT model principles but does not directly use the AT model for topic extraction because, while the AT model supports measuring multi-author research topic distributions within a single paper, it does not consider that scholars with different authorship orders contribute differently to the paper, creating certain limitations and potential biases in the final measurements. Therefore, we analyze author research topic distributions based on the LDA model.

When applying LDA for topic modeling, the number of topics is specified by the modeler. Determining the optimal number of topics is a crucial consideration. This study uses perplexity, a common evaluation metric in statistical language models, to assess model performance and determine the optimal topic number. Perplexity analysis characterizes generative model quality. We employ ten-fold cross-validation (randomly dividing the scholar literature dataset into ten equal parts, using nine parts as training data and one part as test data in each iteration) and use the average of ten perplexity analyses as the final result. A lower perplexity value indicates better model performance. For a test set of M authors, perplexity is defined by Formula (2):

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{m=1}^{M}\ln(P(w_m))}{\sum_{m=1}^{M}N_m}\right\} \quad (2)$$

where D_{test} represents the author test set (a collection of 6,623 author documents), N_m denotes the number of words in the m-th author's corpus, and P(w_m) represents the probability of the LDA model generating that author's text collection, as shown in Formula (3):

$$P(w_m) = \prod p(w_i|z_i = k)p(z_i = k|w_m) \quad (3)$$

X. Wang and A. McCallum empirically verified that JS distance offers greater discriminative power for scholar topic vectors compared to Euclidean distance, cosine distance, and others. Therefore, we adopt JS distance to measure topic similarity between scholars. Let Θ be the set of all discrete probability distributions of author set D over topic set Z. Then for any P, Q Θ, JS distance is calculated by Formula (4):

$$J(P|Q) = \frac{1}{2}\left[\sum p_m \ln \frac{2p_m}{p_m + q_m} + \sum q_m \ln \frac{2q_m}{p_m + q_m}\right] \quad (4)$$

Scholar topic similarity Sim(P, Q) is calculated by Formula (5), where P and Q are the topic distribution vectors of two scholars:

$$\mathrm{Sim}(P, Q) = 1 - JS(P|Q) \quad (5)$$

Feature analysis comprises two parts: (1) correlation analysis between scholar topic similarity and output performance. Using mentor-mentee relationship data from the academic pedigree tree and the obtained topic similarity set, we calculate topic similarity between mentors and students and before/after student graduation. Scholar topic similarity aims to measure the consistency of research directions and fields between two scholars—the degree of knowledge inheritance—and combines this with the biological concepts of "heredity" and "variation" to classify scholars into "hereditary scholars," "non-hereditary non-variation scholars," and "variation scholars." We obtain each scholar's h-index from existing literature as a performance metric and analyze the correlation between knowledge "heredity," knowledge "variation," and academic performance. (2) Research topic distribution of hereditary and variation scholars. We plot the topic distribution of the three scholar types and compare differences in their distributions across topics while conducting a macro-level analysis of the entire pedigree's topic distribution.

2.2 Data Acquisition

Academic pedigree data for this study originates from the project outcomes of the China Association for Science and Technology's "Research on Academic Pedigrees of Contemporary Chinese Scientists," which systematically organized pedigrees in genetics, medicine, chemistry, physics, and agriculture. From this, we identified a clear lineage and intergenerational relationships in genetics, selecting Tan Jiazhen's pedigree—the longest-developed, most comprehensively documented, and largest—as our research object.

Tan Jiazhen graduated from Yenching University Graduate School with a master's degree in 1932 and earned his Ph.D. from the California Institute of Technology in 1936. He established China's first genetics program in the 1950s and taught at Zhejiang University's Biology Department from 1999-2008. During his 60-year teaching career, Tan Jiazhen nurtured numerous academic elites in Chinese genetics, including Ji Daofan, Wang Liquan, Tang Jue, and over 30 others who made outstanding contributions to crop genetics and breeding. Today, Tan Jiazhen's genetics pedigree spans five generations, with detailed mentor-mentee information available for 532 members. This study obtained valid information for 532 scholars within the pedigree, including names, degrees, degree acquisition times, degree-granting institutions, and mentor names. We

define mentor-mentee relationships as intergenerational relationships; the three-generation pedigree tree is shown in Figure 2 [Figure 2: see original paper].

Based on the established pedigree tree, we retrieved all journal literature published by pedigree scholars from CNKI (China National Knowledge Infrastructure). Given the large number of pedigree members, we selected 234 members from the first to third generations (with Tan Jiazhen as the first generation) as our research subjects. Using CNKI, we constructed search queries with the 234 scholars' names, publication dates, subject terms, and author affiliations as elements: (SU='genetics'+'gene'+'DNA'+'chromosome'+'methyl'+'mutation'+'RNA'+'alkyl'+'variation'+'promoter') AND (AU='author name') AND (AF='author affiliation') AND (YEBETWEEN('1960','2017')). We supplemented this with manual verification using scholars' mentor names and degree acquisition times to eliminate false information caused by name duplication, ultimately obtaining 6,623 journal articles (including titles, keywords, abstracts, publication journals, and citation frequencies). Initially, we hoped to analyze SCI papers published by pedigree members. We searched the Web of Science (WOS) database using the advanced query: ((SU=genetics&hereditism) AND Language:(English) AND Document Type:(Article)), retrieving 463,214 genetics papers—an excessively large collection. Given the severe name duplication among Chinese scholars in the SCI database, the absence of full Chinese author names in SCI papers before 2006, and variable institutional name formats, name disambiguation proved difficult, making it challenging to accurately obtain comprehensive information about authors' SCI papers. Meanwhile, the CNKI papers published by pedigree members satisfied our data sample requirements. For these reasons, we selected the CNKI paper collection of pedigree members as our research object.

3 Knowledge Inheritance Analysis of the Genetics Academic Pedigree with Tan Jiazhen as the Core

3.1 LDA Topic Analysis of the Academic Pedigree with Tan Jiazhen as the Core

First, we developed a Python script tailored to the literature dataset characteristics, combining jieba Chinese word segmentation and part-of-speech tagging for data preprocessing. After removing modal particles, adverbs, prepositions, and conjunctions, we obtained over 460,000 feature words from 234 scholars. Using cross-validation to evaluate model performance, we tested topic numbers K = 5, 10, 15, 20, 25, 30, obtaining perplexity values for each case. The results, shown in Figure 3 [Figure 3: see original paper], indicate that perplexity reaches its minimum when the topic number is 20. Therefore, we selected 20 as the optimal topic number for LDA modeling of Tan Jiazhen's genetics pedigree.

Based on the preprocessed dataset, we applied LDA topic modeling to analyze the genetics literature, obtaining topic-word distributions. Due to space limitations, we present only five representative topics among the 20 identified topics,

along with the top 15 words with highest weights in each topic, as shown in Table 1 . By thoroughly understanding genetics research backgrounds, consulting domain experts, and combining this with graduate advisors' research directions, we determined and described the content of each research topic. Based on author-topic distributions, Table 2 shows the research topic distributions of three second-generation pedigree scholars, from which their primary research directions can be observed. For instance, Zhu Lihuang shows the highest distribution value in "Topic 15" (Plant Breeding Genetics), and investigation confirms his research focuses on rice molecular genetics and genomics. Zeng Yitao exhibits the highest value in "Topic 1" (Hematological Genetics), with documented outstanding achievements in globin chemistry, gene structure and function in hemoglobinopathies, and gene therapy for thalassemia. Feng Shuju shows substantial distribution in "Topic 7" (Genomics), and research indicates he studied under Shi Liming, focusing on eukaryotic chromosome structure and function, cytogenetics, and karyotype evolution.

Taking Zhu Lihuang—a representative second-generation scholar with the most students—as an example, we analyzed topic similarity between him and his students. Table 3 lists the top five and bottom five similarity values between Zhu Lihuang and his students. Calculations for all data yielded similarity values ranging from [0.487, 0.747] for pedigree members.

3.2 Intergenerational Knowledge Inheritance in the Academic Pedigree with Tan Jiazhen as the Core

In biological theory, heredity refers to the phenomenon where offspring repeat parental characteristics (traits) in a continuous system, essentially involving offspring inheriting parental genetic material—genes (which determine biological traits). Gene transmission constitutes heredity. During generational continuity, gene mutations enable offspring to develop traits different from their parents, a phenomenon known as variation. Knowledge is considered the "gene" transmitted from mentors (parents) to students (offspring) through oral instruction, face-to-face teaching, and various other methods. Drawing upon biological concepts of heredity and variation, we view the continuation of research directions between mentors and students as knowledge "heredity" within the pedigree, and students' research direction shifts during their subsequent academic careers as knowledge "variation."

Topic similarity exhibits continuous value ranges. To facilitate identification of correlations between mentor-mentee knowledge inheritance and academic performance, we adopt the statistical concept of interquartile range (IQR), also known as the midspread. As a method in descriptive statistics, IQR determines the difference between the third and first quartile intervals. Like variance and standard deviation, IQR indicates the dispersion of variables in statistical data, but it is more robust. Therefore, using IQR to identify and partition "heredity" and "variation" in knowledge inheritance is statistically meaningful. We define the top quarter of the total sample's topic similarity range as the "variation

interval" in knowledge inheritance, with scholars in this interval classified as "variation scholars." The bottom quarter is defined as the "heredity interval," with scholars in this interval classified as "hereditary scholars" (in practice, heredity and variation in knowledge inheritance are relative: hereditary scholars exhibit some research topic variation but to a lesser degree, while variation scholars show some topic heredity but to a lesser extent). Using this method, we calculated the heredity and variation topic similarity intervals, as shown in Figure 4 [Figure 4: see original paper].

The figure shows that within Zhu Lihuang's academic pedigree, two members exhibit research topic heredity, while seven show research direction variation. Overall, 21 members (8.97%) demonstrate research topic heredity, and 34 members (14.53%) show research direction variation. The topic similarities between these 54 scholars and their mentors are presented in Table 4 . The pedigree tree contains 22 academic branches with Tan Jiazhen and his direct students at the core. The distribution quantities and proportions of the three scholar types across different branches are shown in Figure 5 [Figure 5: see original paper].

Generally, students' pre-graduation research follows their mentors' footsteps, so their research topics typically align with their mentors before degree completion. Variation clearly arises from post-graduation research topic changes. Therefore, we use the second preprocessed sample type as input for the LDA topic model. The specific process is as follows: (1) Divide each scholar's literature into pre-graduation and post-graduation publication sets based on graduation and publication dates; (2) Treat scholar i's pre-graduation and post-graduation literature sets as two distinct entities A_i and B_i, using the LDA model to obtain their corresponding topic distribution vectors P_a and P_b; (3) Calculate the JS distance between P_a and P_b to determine scholar i's topic similarity before and after graduation. As shown in Figure 6 [Figure 6: see original paper], the topic similarity change interval for students before and after graduation is [0.504, 0.780], with heredity and variation intervals of [0.711, 0.780] and [0.504, 0.573], respectively.

Similarly, we calculated that the proportions of scholars exhibiting heredity and variation in research topics before and after graduation are 9.66% and 24.14%, respectively. There are 14 "hereditary scholars" and 35 "variation scholars." Clearly, compared with intergenerational academic inheritance results, the proportion of "hereditary scholars" remains basically unchanged, while the proportion of "variation scholars" increases significantly when examining changes in students' own research topics before and after graduation.

4 Correlation Analysis Between Topic Similarity and Academic Performance in the Academic Pedigree with Tan Jiazhen as the Core

This study seeks to determine whether knowledge inheritance, represented by intrinsic textual connections, influences scholars' career development and, if so,

how and to what extent. Therefore, we conducted correlation analysis between the degree of knowledge inheritance among pedigree members and their individual research performance, using the h-index as the performance evaluation metric.

As shown in Figure 7 [Figure 7: see original paper], panels (1) and (2) display the distribution of intergenerational topic similarity versus h-index for scholars in Tan Jiazhen's pedigree, while panels (3) and (4) show the distribution of pre/post-graduation topic similarity versus h-index. For analytical convenience, we designate the intergenerational topic similarity and corresponding h-index data as Combination 1, and pre/post-graduation topic similarity and corresponding h-index data as Combination 2. Different colored areas represent "variation scholars" (Variation), "non-hereditary non-variation scholars" (Non-hereditary Non-variation), and "hereditary scholars" (Heredity).

Panels (1) and (3) show scatter distributions for Combinations 1 and 2, respectively. In panel (1), point distribution exhibits clustering characteristics, with 71% of scholars' similarity values falling within the narrow interval (0.55, 0.65). The proportions of "hereditary scholars," "non-hereditary non-variation scholars," and "variation scholars" are 8.97%, 76.50%, and 14.53%, respectively, with mean h-index values of 5.42, 4.54, and 7.10. Although "hereditary scholars" and "variation scholars" represent smaller proportions, they demonstrate higher academic performance than "non-hereditary non-variation scholars." The proportions of high-performance authors (h-index > 5) in the three intervals are 3.00%, 56.92%, and 35.38%, respectively, indicating that "variation scholars" have a relatively higher proportion of high-performance individuals. Similarly, in panel (2), the proportions of the three scholar types are 9.66%, 66.20%, and 24.14%, with mean h-index values of 5.42, 5.52, and 5.11. High-performance author proportions are 8.00%, 76.00%, and 16.00%, respectively, showing that "hereditary scholars" and "variation scholars" constitute higher proportions among high-performance researchers than in the overall pedigree population.

To further investigate potential correlations between topic similarity and h-index, we divided all scholars into 30 equal intervals based on topic similarity. For each interval, we plotted a point where the x-coordinate represents the mean h-index of that interval and the y-coordinate represents the interval's midpoint. After processing both datasets using this method, we obtained scatter plots shown in panels (2) and (4) of Figure 7. Curve fitting was further applied to these scatter plots. The blue solid lines represent fourth-degree polynomial fitting results:

$$y_1 = -12823.5x^4 + 18522.5x^3 - 9533.4x^2 + 2050.1x - 147.8$$

$$y_2 = 4934.2x^4 - 7555.2x^3 + 4201.7x^2 - 1005.8x + 90.3$$

The R^2 values are 0.699 for y_1 and 0.068 for y_2. R^2 measures goodness-of-fit for least squares curve fitting, with larger values indicating better fit (ranging

from 0 to 1). Clearly, y_1 shows good fit. Panel (2) reveals a flat concave distribution after curve fitting—low in the middle and high on both sides. This concave distribution indicates that scholars with intergenerational topic similarity in the "heredity" and "variation" intervals exhibit significantly higher academic performance than "non-hereditary non-variation" scholars, suggesting that high-performance researchers are not the most numerous "non-hereditary non-variation" group but rather the minority groups of hereditary and variation scholars. Panel (4) shows that the fitted mean h-index remains relatively stable (4-8) across pre/post-graduation topic similarity changes.

The above intergenerational topic similarity reveals overall characteristics of research topic changes between students and mentors but does not address whether students' post-graduation research directions differ from their mentors'. Therefore, we further calculated similarity between students' post-graduation research topics and their mentors' topics. We selected students' post-graduation literature sets and mentors' literature sets for topic extraction and similarity calculation, then analyzed the correlation between topic similarity and h-index. The results are shown in panels (5) and (6) of Figure 7. The similarity range between students' post-graduation topics and mentors' topics is (0.27, 0.74), with an expanded range due to the removal of duplicate texts (co-authored papers during graduate studies). The proportions of "hereditary scholars," "non-hereditary non-variation scholars," and "variation scholars" are 6.20%, 68.21%, and 25.58%, respectively, with mean h-index values of 6.62, 4.81, and 7.24. High-performance author proportions in the three intervals are 10.25%, 57.69%, and 32.05%, respectively. Similar to Combination 1, "hereditary scholars" and "variation scholars" constitute higher proportions among high-performance researchers than in the overall population. The fitted curve again shows a flat concave distribution, confirming that knowledge "heredity" and "variation" contribute to improved academic performance:

$$y_3 = -3762.8x^4 + 6857.6x^3 - 4464.7x^2 + 1223.1x - 112.73$$

The R^2 value for y_3 is 0.70.

5 Research Topic Distribution of Hereditary and Variation Scholars in Tan Jiazhen's Academic Pedigree

The above research demonstrates that "hereditary scholars" and "variation scholars" achieve relatively high academic performance during their career development. We also developed strong interest in the research topics of these two scholar types: What are their topic distributions? Do significant differences exist? To address these questions, we further employed topic modeling to analyze the research topic distributions of "hereditary scholars" and "variation scholars."

The maximum likelihood estimate of obtained from the LDA topic model represents each scholar's topic distribution. Extending scholar document collections

to the scholar group within the pedigree, we can similarly calculate the group's topic distribution . Let _d^k be the proportion of topic k for scholar d. Then the intensity _s^k of scholar group s in topic k is given by Formula (6), with values between 0 and 1:

$$\theta_s^k = \frac{1}{|s|}\sum_{d=1}^{|s|}\theta_d^k \quad (6)$$

Using each topic's _s^k to represent the intensity of topic k in target scholar group s, we can derive the topic distribution of the target pedigree population, as shown in Figure 8 [Figure 8: see original paper]. Blue represents "hereditary scholars" identified through intergenerational topic similarity calculations, green represents "variation scholars," and red represents the entire scholar population of the pedigree. The figure reveals: (1) Topic 14 (Cytogenetics) shows the highest intensity across all three groups, with intensities of 0.53, 0.27, and 0.33 for "hereditary scholars," "variation scholars," and "all scholars," respectively. Consultation with experts confirms this is fundamental genetics research and a prerequisite for other studies. (2) "Hereditary scholars" primarily concentrate in Topic 12 (Animal Genetics), Topic 14 (Cytogenetics), Topic 17 (Tumor Cytogenetics), and Topic 20 (Population Genetics), with minimal distribution in other topics. (3) "All scholars" and "hereditary scholars" show more dispersed and relatively uniform topic distributions compared to "hereditary scholars." (4) Some "variation scholars" devote themselves to plant genomics, medical genetics, and anti-toxin application research—areas where "hereditary scholars" are scarcely involved. (5) Some "hereditary scholars" focus on animal genetics research, a field where "variation scholars" rarely conduct studies.

6 Conclusions and Future Work

This study employed LDA topic modeling to investigate topic distributions within Tan Jiazhen's genetics academic pedigree, exploring pedigree members' research topic distributions at the semantic level. By calculating research topic similarity, we classified scholars into "hereditary scholars," "non-hereditary non-variation scholars," and "variation scholars," and further analyzed correlations between topic similarity and research performance. The results indicate: (1) Scholars' research performance correlates with the degree of intergenerational topic change but not with the degree of topic change before and after graduation. (2) The average research performance of "hereditary scholars" and "variation scholars" significantly exceeds that of "non-hereditary non-variation scholars," suggesting that moderate "respecting and continuing predecessors' work" and "pioneering new territories" both enhance research performance. (3) "Hereditary scholars" and "variation scholars" exhibit substantially different research topic distributions. Overall, cytogenetics remains a fundamental research area in genetics, providing necessary knowledge accumulation for all three scholar types.

This study effectively extracts scholars' research topics at the semantic level and measures intergenerational topic similarity within Tan Jiazhen's pedigree using JS distance, improving research credibility compared to previous qualitative studies. Research on talent development pathways based on research topics helps reveal talent growth patterns and inform talent policy formulation. Research evaluation systems and incentive mechanisms should fully consider the "hereditary genes" of knowledge inheritance, providing longer research cycles for researchers who continue in the same direction to promote deep scientific innovation based on solid research foundations. Conversely, they should encourage researchers to explore emerging or interdisciplinary fields on the basis of knowledge inheritance, offering financial support or resources to facilitate greater progress in knowledge transformation.

This research is limited to knowledge inheritance in the genetics academic pedigree, and the results have domain-specific limitations. Future work will consider disciplinary differences and explore knowledge inheritance in other research fields. Additionally, since characteristic differences between Chinese and foreign scientific research fields are evident, we will further investigate knowledge inheritance features in foreign academic pedigrees.

References

[1] Liu Ying, Zhang Yanlei, Zhang Daqing. Preliminary exploration and practice of constructing a Chinese scientists' academic pedigree database [J]. Library and Information Service, 2014, 58(S2): 60-62.

[2] Cronin B, Sugimoto C. Academic genealogy [C] // Blaise C, Cassidy R. Beyond bibliometrics. London: MIT Press, 2014.

[3] Jackson D C. Academic genealogy and direct calorimetry: a personal account [J]. Advances in physiology education, 2011, 35(2): 120-128.

[4] Malmgren R D, Ottino J M, Amaral L A N. The role of mentorship in protégé performance [J]. Nature, 2010, 465(7298): 622-626.

[5] Chang Huan, Lü Ruihua, Zhang Jiajing. Research on collaboration networks within academic pedigrees: A case study of the Quaternary academic pedigree with Liu Dongsheng as the core [J]. Information Studies: Theory & Application, 2016, 39(4): 14-19.

[6] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003, 3(3): 993-1022.

[7] Rosen Z, Michal, Griffiths T, et al. The author-topic model for authors and documents [C] // Proceedings of the 20th conference on Uncertainty in artificial intelligence. New York: AUAI, 2004: 487-494.

[8] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering [J]. Machine learning, 2001, 42(1): 143-175.

[9] Shi Qingwei, Qiao Xiaodong, Xu Shuo, et al. Author-topic evolution model and its application in research interest evolution analysis [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(9): 912-919.

[10] Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends [C] // Backstrom L, Huttenlocher D, Kleinberg J, et al. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2006: 424-433.

[11] Xie Ping. The origin of life—abandonment and innovation of evolutionary theory [M]. Beijing: Science Press, 2014.

[12] Liu Junwan, Zheng Xiaomin, Su Na, et al. Quantitative analysis of publication delay in domestic and foreign informetrics journals: A case study of Scientometrics and Journal of the China Society for Scientific and Technical Information [J]. Chinese Journal of Scientific and Technical Periodicals, 2016, 27(12): 1292-1299.

[13] Cui Kai. Research and implementation of topic evolution based on LDA [D]. Changsha: National University of Defense Technology, 2010.

Author Contributions:
Liu Junwan: Research conceptualization and design, data analysis, manuscript writing;
Yang Bo: Data collection and cleaning, program code design, data analysis, manuscript writing;
Wang Feifei: Research design, manuscript revision;
Xu Shuo: Topic modeling methodology guidance.

Note: Figure translations are in progress. See original paper for figures.