國立臺灣大學電機資訊學院暨中央研究院

資料科學學位學程

碩士論文

Data Science Degree Program

College of Electrical Engineering and Computer Science

National Taiwan University and Academia Sinica

Master's Thesis

科學合作網絡的綜合研究:基於網絡分析技術

A comprehensive study on the scientific collaboration networks via techniques in network analysis

賴銘彥

Ming-Yen Lai

指導教授：潘建興 博士、楊鈞澔 博士

Advisor: Kin-Hing Phoa, Ph.D. , Chun-Hao YANG, Ph.D.

中華民國 114 年 07 月

July 2025

doi:10.6342/NTU202501640

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

科學合作網絡的綜合研究：基於網絡分析技術

A comprehensive study on the scientific collaboration networks via techniques in network analysis

本論文係 賴銘彥（R10946016）在國立臺灣大學資料科學學位學程完成之碩士學位論文，於民國 112 年 6 月 26 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Data Science Degree Program on 26 June 2023 have examined a Master's thesis entitled above presented by MING-YAN LAI (R10946016) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

（指導教授 Advisor）	（指導教授 Advisor）	

學程主任 Director:

致謝

在這篇論文完成之際，我要向所有在求學路上給予我支持與幫助的人們表達最深切的感謝。

首先，我要感謝我的家人。沒有你們無條件的愛與支持，我不可能在學術路上如此任性地追求自己的理想。你們的理解與包容，讓我能夠專心致志地投入研究工作，即使在最困難的時刻也不曾放棄。你們是我最堅強的後盾，也是我前進的動力。

其次，我要向我的指導教授 Fred 表達最誠摯的謝意。我非常有幸能夠在您的指導下完成這項研究。您淵博的學識、嚴謹的治學態度，以及耐心的指導，不僅讓我在學術上獲得成長，更重要的是培養了我獨立思考和解決問題的能力。您總是能夠在我迷茫時給予方向，在我遇到瓶頸時提供關鍵的建議，這份恩情我將銘記在心。

同時，我要特別感謝我的女朋友 Jennifer。很多有趣的靈感都來自於與你的討論和交流。你的聰慧是多領域且多模態的，總能從不同的角度為我帶來新的思考視野。在我研究遇到困難時，你總是耐心地聽我傾訴，給予我鼓勵和支持。你的存在不僅豐富了我的學術思考，更讓我的求學生活充滿了溫暖與快樂。

此外，我也要感謝所有在研究過程中給予我協助的老師、同學和朋友們。感謝實驗室的夥伴們與我分享經驗，互相討論學術問題；感謝圖書館的工作人員提供便利的研究環境；感謝所有曾經給予我建議和幫助的人們。

最後，謹以這篇論文獻給所有支持我、相信我的人。這份成果屬於我們大家。

謝謝大家。

2025.07.21 賴銘彥

doi:10.6342/NTU202501640

中文摘要

隨著研究主題變得專業化和多元化，政府和組織如何有效分配有限的研究資源已變得至關重要但極具挑戰性。馬太效應是指成功的作者往往會更加成功，而知名度較低的作者則往往難以獲得認可的現象，這種現象在科學合作網絡中被觀察到。在我們之前的研究中，我們在不同假設條件下建立了科學合作網絡模型，並量化了流行效應（也稱為馬太效應[15]）對研究人員和學科的影響，以及研究人員在該學科中的天才程度。在本文中，我們將此模型應用於從科學網（WoS）資料庫收集的不同學科，並根據模型中每個學科的參數對其進行分析。我們使用這些指標創建了一個圖表，並分析了任意兩個學科之間這兩個指標相似性的關係

關鍵字:。馬太效應, Web of Science, 合作網絡, 相關度, 流行效應

doi:10.6342/NTU202501640

Abstract

As research topics become specialized and diverse, it has become crucial but challenging for governments and organizations to distribute limited research resources effectively. The Matthew effect, which refers to a phenomenon where successful authors tend to be more successful while lesser-known authors tend to struggle to gain recognition, is observed in the scientific collaboration network. In our prior research, we developed a scientific collaboration network model under different hypotheses and quantified the impact of the popular effect (also known as the Matthew effect[15]) on both the researcher and the subject, as well as the researcher's level of genius within the subject. In this paper, we apply this model to different subjects collected from the Web of Science (WoS) database and analyze them based on the parameters of each subject in the model. We used these indicators to create a plot and analyze the relationship between the similarities of the two indicators for any two subjects.

Keywords: Popular effect, Web of Science, Rank of relation, Co-authorship network

doi:10.6342/NTU202501640

目　次

doi:10.6342/NTU202501640

1 Introduction

Scientific collaboration between different countries and research institutions has become increasingly common, not only because different cultures can bring different research ideas, but even more funding for research or citation.

According to the network analysis, the scientific collaboration network is discussed in various aspects, including research fields and regions[1], publishing habits[16], international cooperation or not, gender[7], etc., and most importantly, the popular effect.[17]

Popularity effect, which states that the strong get stronger and the weak get weaker, was first proposed in 1968 and is commonly observed in the real world [9][15]. In scientific collaboration network, we believe that a similar popularity effect is present for three main reasons. First, popular authors receive more attention and have more opportunities to collaborate with others. Second, popular authors have more students and researchers working with them to co-author papers. Third, journal reviewers tend to favor papers written by popular authors.

Therefore, we propose a leading author model to analyze the degree of popularity effect in scientific collaboration networks. One notable aspect of our model compared to others is that we assume that the first author does most of the work on a paper, rather than the work being divided equally among all authors. This assumption is more in line with the reality of scientific collaboration networks.[18]

In the previous paper[14], we applied this model to try to find out who are the truly influential nodes[13] in a network during a certain period of time. Because the model takes into account the popular effect, the nodes we found are more objective and closer to reality (that is, those who are not easily co-authored because of fame, but because of the high productivity of the nodes themselves).

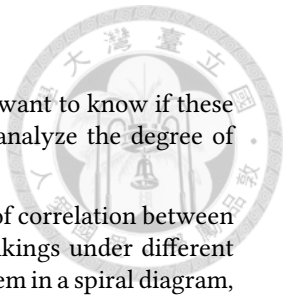
This paper focus on applying the leading author model to each subject in the Web of Science(WOS) database, with four coefficient that we care about in the model.In an attempt to find the correlation between different subjects

1. $\hat{\theta_l}$: controls the total number of papers per unit time.
2. $\hat{\theta_g}$: means the genius coefficient which determines the size of the effect of authors' genius levels. The authors with high genius levels will contribute more to the scientific collaborative system.
3. $\hat{\theta_c}$: is related to the number of paper's authors.
4. $\hat{\theta_u}$: The popularity coefficient represents the magnitude of the popularity effect. A positive u indicates that the rich-get- richer effect is working in the collaborative system, whereas a negative u shows the opposite direction, namely the rich-get-poorer.

First, we estimate the four coefficient and calculate the standard deviation for the 182 subjects in WoS.

Second, according to different coefficients, we perform a t-test on any two subjects among 182 subjects, and sort them based on the size of the t-statistic.

doi:10.6342/NTU202501640



Given the rankings of t-statistic for subject pairs under different coefficients, we want to know if these rankings have consistency. Therefore, we use Spearman's rank correlation to analyze the degree of correlation between any two coefficients.

Finally, we hope to use the rankings of these 4 parameters to measure the degree of correlation between different subjects (i.e. if two subjects are ranked very high in the t-statistic rankings under different parameters, it means that these two subjects are more similar). And we will plot them in a spiral diagram, with the center point being the subject we want and the other 181 subjects plotted as points around it. The closer the point is to the center, the more correlated it is.

2 A Literature Review in the Field of Scientometrics

Prior Studies

In the field of scientometrics, there are several major directions:

1. Academic impact assessment
2. Scientific collaboration and co-authorship networks
3. Academic innovation and scientific discovery
4. Science evaluation and science policy
5. Science communication and science indicators

Understanding the structure and functioning of scientific collaboration networks is crucial for fostering interdisciplinary collaboration, enhancing research efficiency, and driving scientific progress. In my research, I have employed The Leading Author Model to investigate author collaboration networks across different disciplines.

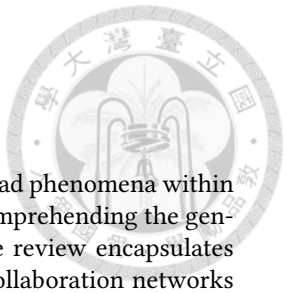
The following four studies related to this paper are enumerated.

Relevant Papers 1.Evolution of networks with aging of sites

"Evolution of networks with aging of sites"[5] by S. N. Dorogovtsev and J. F. F. Mendes, which investigates the impact of network node aging on network growth and evolution. This study introduces the concept of node aging, defined as the decrease in a node's ability to attract new connections over time, implemented through a probability function related to the node's age. The findings indicate that considering node aging, the network's connectivity distribution and growth dynamics exhibit unique scaling behavior, significantly differing from the traditional Barabási-Albert model.

In my study of academic collaboration networks, the concept of node aging aids in understanding the Matthew effect, where resources and recognition tend to concentrate around scholars or institutions already possessing high prestige. Over time, some scholars may become more active due to the accumulation of achievements, while others may become marginalized due to a lack of resources. Analyzing the rate of node aging and network evolution reveals how the Matthew effect can be reinforced or mitigated through changes in network structure.

doi:10.6342/NTU202501640



Relevant Papers 2.Science of science

The review paper "Science of Science" plays a pivotal role in delving into the myriad phenomena within scientific research, particularly highlighting the utility of big data analytics in comprehending the genesis, dissemination, and evolution of scientific knowledge. This comprehensive review encapsulates various dimensions of scientific inquiry, including the intricacies of academic collaboration networks and the manifestation of the Matthew effect within the realm of research activities.

Moreover, "Science of Science" leverages an extensive analysis of metadata from a plethora of academic publications and collaborative endeavors, thereby uncovering patterns and trends prevalent in the scientific research landscape. This analysis furnishes my study with quantitative methodologies and tools essential for dissecting and quantifying the influence of the Matthew effect within academic collaboration networks. The arsenal of methods employed, ranging from network analysis to statistical modeling and the theories of complex systems, serves as the cornerstone for deciphering and elucidating the data pertinent to my research.

Relevant Papers 3.Why do papers from international collaborations get more citations? A bibliometric analysis of Library and Information Science papers?

This paper, through bibliometric analysis, investigates the relationship between international collaboration and the citation count of scholarly articles, revealing that international collaborations often significantly increase citation rates. This discovery holds substantial importance for my research, as it directly ties to the Matthew effect within academic collaboration networks—where researchers already recognized for their visibility and accolades can further amplify their academic influence through international collaborations. This phenomenon could exacerbate the existing inequalities within academia, where resources and recognition are increasingly concentrated among a select few top scholars or institutions.

My research area focuses on academic collaboration networks and their susceptibility to the Matthew effect, especially in terms of how collaborations can enhance the impact and visibility of research findings. Against this backdrop, the paper "Why do papers from international collaborations get more citations? A bibliometric analysis of Library and Information Science papers" offers a crucial insight into the potential impact of international collaborations on the citation frequency of academic papers.

Relevant Papers 4.Higher-order rich-club phenomenon in collaborative research grant networks

The paper "Higher-order rich-club phenomenon in collaborative research grant networks" unveils the existence of a higher-order rich-club phenomenon within research funding networks, providing concrete evidence that a specific elite group within academia tends to form tight-knit collaboration circles, exacerbating the unequal distribution of academic resources.

This finding offers a tangible case study for my research, illustrating how the Matthew effect is manifested and intensified in academic collaboration networks through the mechanisms of research funding allocation.

These insights enrich my understanding of the dynamics within academic networks and highlight the need for strategies to promote a more equitable distribution of resources, aiming to foster a more inclusive and diverse academic environment.

doi:10.6342/NTU202501640

3 The leading author model

The leading author model is a new and better model to explain the scientific collaboration than the standard network analysis.

Main concept Suppose we have time-series collaboration data at time t = 0,1,...,T. Let V_t be the author shows in the data at time t = 0,1,...,T. We mark that an author i join and leaves the system at time $t_{0,i}$ and $t_{1,i}$, respectively. So we can say $i \in V_t$ if and only if $t = t_{0,i}, t_{0,i} + 1, ..., t_{1,i}$. For each author i in time t, has popularity $u_{i,t}$,

which $u_{i,t} = \frac{1}{t-1+T_0}$ (number of papers of author i until time t -1) . (1)

where T_0 is the duration of the initial data at t=0. and set g_i, the genius level of author i in the system. Assumed that g_i follows the standard normal distribution. So the p.d.f of g_i can be written as $p = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}g_i^2)$, $-\infty < g_i < \infty$. (2)

We assumed two more things, first, there is a leading author who lead the whole paper's work and gather co-authors in every paper. And second is strong assumption that if a author is popular, he has more chance to become a co-author of a paper,chosen by the leading author.

The following is the paper generation process

-For t = 1,2,...,T :

-For $i \in V_t$:

-Draw the number of leading-authored papers of author i at time t:

$$d_{i,t} \sim Poisson\,(\exp(\theta_l + \theta_g\, g_i\,)) \tag{3}$$

-For each paper led by author i, d =1,2,...,$d_{i,t}$:

-Choose co-authors of paper $q_{i,t,d}$ among $j \in V_{i,t}$ with probability:

$$P(\text{ j becomes a coauthor of } q_{i,t,d}) = S(\,\theta_c + \gamma\,\theta_g\theta_j + \theta_u u_{j,t}) \tag{4}$$

where $S(x)=\frac{1}{1+e^{-x}}$ is a sigmoid function.

For an author, his genius level g_i affects writing of both leading-authored and co-authored papers, and his popularity $u_{j,t}$ affects only writing of co-authored papers. The popularity coefficient θ_u represents the magnitude of the popularity effect. If θ_u is positive, it means that the rich-get-richer effect is working in the collaborative system, on the contrary, a negative θ_u shows the opposite direction.

The next parameter θ_g, genius coefficient, determines the size of the effect of authors' genius levels. From the function above, we know that the authors with high genius levels will produce more to the network. The level of contribution can be different between leading-authored and co-authored papers, which is controlled by the positive-valued hyperparameter γ. If γ is samll, then the contribution is mainly measured by the leading-authored papers. In this paper, we use γ = 1

4 Algorithm

4.1 Preliminary

doi:10.6342/NTU202501640

In this section, we present an algorithm for estimating the model parameters, the genius levels of the authors, and the leading authors of each paper. The EM algorithm will be used in conjunction with the Gibbs sampling method to alternately update these parameters. We assume that the hyperparameter γ is given.

We denote by $c_t = \sum_{i \in V_t} d_{i,t}$ the total number of papers written at time t, and $m_{i,t}$ presents the number of co-authored(non-leading) papers of author i at time t.

Based on the generative process described above, an author i can join c_t-$d_{i,t}$ papers with a probability of $S(\theta_c + \gamma\theta_g g_i + \theta_u u_{i,t})$. Therefore, we have

$$m_{i,t} \sim Binomial(c_t\text{-}d_{i,t},\ \theta_c + \gamma\theta_g g_i + \theta_u u_{i,t}) \tag{5}$$

And the probability mass function is given by

$$p(m_{i,t}|c_t, d_{i,t}, g_i, u_{i,t}, \theta) = \left(\frac{c_t - d_{i,t}}{m_{i,t}}\right)(S(\theta_c + \gamma\theta_g g_i + \theta_u u_{i,t}))^{m_{i,t}}(1 - S(\theta_c + \gamma\theta_g g_i + \theta_u u_{i,t}))^{c_t - d_{i,t} - m_{i,t}} \tag{6}$$

for $m_{i,t}$=0,1,...,c_t-$d_{i,t}$, which can be simplified to

$$p(m_{i,t}|c_t, d_{i,t}, g_i, u_{i,t}, \theta) = \left(\frac{c_t - d_{i,t}}{m_{i,t}}\right)exp(m_{i,t}(\theta_c + \gamma\theta_g g_i + \theta_u u_{i,t})) \times (1 + exp(m_{i,t}(\theta_c + \gamma\theta_g g_i + \theta_u u_{i,t})))^{-c_t + d_{i,t}}. \tag{7}$$

The probability mass function of $d_{i,t}$ can be expressed by

$$p(d_{i,t}|g_{i,\theta}) = \frac{1}{d_{i,t}!}exp(\theta_l + \theta_g g_i)exp(\text{-}exp(\theta_l + \theta_g g_i)),\ d_{i,t} = 0, 1, \ldots \tag{8}$$

Note that $d_{i,t} + m_{i,t}$ is the total number of papers written by author i at time t.

For a paper $q_{i,t,d}$, which is the d-th leading paper of author i at time t , we aim to identify the leading author i among all the observed authors of the paper. We reindex the papers at time t, from $q_{i,t,d}$, $i \in V_t$, d = 1,..., $d_{i,t}$ to $q_{c,t}$, c = 1,...,c_t. Let $b_{c,t} \in V_t$ be the leading author of paper $q_{c,t}$.

For simplicity, we denote the set of variables by omitting the subscripts t, i, c, and so on. For example,

g =$\{ g_i : i \in V\}$, $u_i = \{u_{i,t}: t = \max(t_{0,i}, 1),...,t_{1,i}\}$, $u_t = \{u_{i,t} : i \in V_t, \}$,$\mathbf{u} = \{u_{i,t} : i \in V, t = \max(t_{0,i},1),...,t_{1,i}\}$, $b_t = \{b_{c,t} : c = 1,...,c_t\}$, $\mathbf{b} = \{b_{c,t} : t = 1,...,\mathbf{T}, c = 1,...,c_t\}$, and $\mathbf{c} = \{c_t : t = 1,..., \mathbf{T}\}$, where $\mathbf{V} = \bigcup_{t=1}^{T} V_t$. We use max ($t_{0,i}$,1) for $t_{0,i} = 0$ since the stochastic inference will be made on $t = 1,...T$.

4.2 The Likelihood Function

In this part, we explain distribution functions of the model.We identify the three parts of the generative process: genius levels, leading-authored papers, and co- authored papers. First, the probability distribution function for the authors' genius levels can be written by

$$p(\mathrm{g}) = \prod_{i \in V} p(g_i). \tag{9}$$

Second, the probability distribution function of the number of leading au-thored papers is expressed by

$$p(\mathbf{d} \mid \mathbf{g}, \theta) = \prod_{t=1}^{T} \prod_{i \in V_t} p(d_{i,t}|g_i, \theta). \tag{10}$$

Third, the remaining process is about the selection of co-authors of papers, and the probability distribution function is given by

$$p(\mathbf{m} \mid \mathbf{d}, \mathbf{g}, \mathbf{u}\theta) = \prod_{t=1}^{T} \prod_{i \in V_t} p(m_{i,t}|c_t, d_{i,t}, g_i u_{i,t}, \theta). \tag{11}$$

doi:10.6342/NTU202501640

Combining the three processes, we obtain the total probability distribution function of our generative stochastic model, given by

$$p(\mathbf{g}, \mathbf{d}, \mathbf{m} \mid \mathbf{u}, \theta) = p(g)p(\mathbf{d} \mid \mathbf{g}, \theta)p(\mathbf{m} \mid \mathbf{d}, \mathbf{g}, \mathbf{u}, \theta) \qquad (12)$$

using the Bayes Theorem. We then have the complete data likelihood function of the model, given by $L(\theta \mid \mathbf{g}, \mathbf{d}, \mathbf{m}, \mathbf{u}) = p(\mathbf{g},\mathbf{d},\mathbf{m}| \mathbf{u}, \theta)$. The complete data log-likelihood function for the EM algorithm can be obtained by

$$l(\theta \mid \mathbf{g}, \mathbf{d}, \mathbf{m}, \mathbf{u}) = lnp(\mathbf{g}, \mathbf{d}, \mathbf{m} \mid \mathbf{u}, \theta) = lnp(g) + lnp(d|\mathrm{g},\theta) + lnp(\mathbf{m} \mid \mathbf{d}, \mathbf{g}, \mathbf{u}, \theta) \qquad (13).$$

For simplicity, we rewrite the log-likelihood function by $l(\theta \mid \mathbf{g}, \mathbf{b}, \mathbf{m}, \mathbf{u}) = l(\theta \mid \mathbf{g}, \mathbf{d}, \mathbf{m}, \mathbf{u})$

since we can determine **d** and m by the authors list **q** and the leading author information **b** of all the papers at t = 1,...,T. Now we de ne the expected value of the complete data log-likelihood function for the model as

$$Q(\theta \mid \hat{\theta}^{(s)}) = E\left[l(\theta \mid \mathbf{G}, \mathbf{B}, \mathbf{q}, \mathbf{u}) \mid \mathbf{q}, \mathbf{u}, \hat{\theta}^{(s)}\right] \qquad (14)$$

where the expectation is taken over the uppercase quantities **G** and **B** that correspond to the genius levels **g** of authors and the choices of the leading author **b**, respectively. Here, $\hat{\theta}^{(s)}$ is the parameter estimate at the s-th iteration of the EM algorithm. This function is to be maximized in the M-step of the EM algorithm.

4.3 Gibbs Sampling

Computing $Q(\theta \mid \hat{\theta}^{(s)})$ can be a difficult task since it usually cannot be expressed in a closed form. To overcome this issue, a possible solution is to use a sampling approach.

Rather than computing the analytical solution, we will adopt a sampling approach and draw samples of **G** and **B** from the distribution $p(\mathbf{g}, \mathbf{b} \mid \mathbf{q}, \mathbf{u}, \hat{\theta}^{(s)})$. Using these samples, we can estimate the expected value over **G** and **B**.

In order to sample authors' genius levels using the Gibbs sampling technique, we will alternate between sampling **G** and **B** from their respective conditional distributions. Specifically, the following conditional distributions of **G** will be used as target distributions for sampling the authors' genius levels:

$$p\left(g_i \mid \mathbf{g} \backslash \{g_i\}, \mathbf{b}, \mathbf{q}, \mathbf{u}, \widehat{\theta}^{(s)}\right) \propto p\left(g_i\right) \prod_{t=\max(t_{0,i},1)}^{t_{1,i}} p\left(d_{i,t} \mid g_i, \widehat{\theta}^{(s)}\right) p\left(m_{i,t} \mid c_t, d_{i,t}, g_i, u_{i,t}, \widehat{\theta}^{(s)}\right) (15)$$

For $i \in V$. Note that $g_j, j \neq i$ are not conditioned eventually since they are independent to the conditional distribution of g_i. The above relationship can be easily derived from Bayes' theorem.

Next, we will proceed with the sampling of **B**, which accounts for choosing a leading author among all the paper authors. For each paper $q_{c,t}, t = 1, \cdots, T, c = 1, \cdots, c_t$, we use Bayes' theorem to calculate the following:

$$p\left(b_{c,t} = i_k \mid \mathbf{g}, \mathbf{b} \backslash \{b_{c,t}\}, \mathbf{q}, \mathbf{u}, \hat{\theta}^{(s)}\right) = p\left(b_{c,t} = i_k \mid \text{authors of } q_{c,t} \text{ is } \{i_1, \cdots, i_K\}, \mathbf{g}, \mathbf{u}, \hat{\theta}^{(s)}\right) \propto$$
$$p\left(b_{c,t} = i_k \mid \mathbf{g}, \mathbf{u}, \widehat{\theta}^{(s)}\right) p\left(\{i_1, \cdots, i_K\} \setminus \{i_k\} \text{ are coauthors of } q_{c,t} \mid b_{c,t} = i_k, \mathbf{g}, \mathbf{u}, \widehat{\theta}^{(s)}\right), \qquad (16)$$

where $\{i_1, \cdots, i_K\}$ is the set of authors of paper $q_{c,t}$. Our goal is to calculate the probability

doi:10.6342/NTU202501640

that i_k is the leading author. It should be noted that $p\left(b_{c,t}=i_k \mid \mathbf{g},\mathbf{u},\widehat{\theta}^{(\mathbf{s})}\right)$ is the probability for i_k to become a leading-author of any paper among all authors at time t. Since the number of leading-authored papers for author i is determined by the Poisson distribution with mean parameter $\exp\left(\theta_l+\theta_g g_i\right)$, we have

$$p\left(b_{c,t}=i_k \mid \mathbf{g},\mathbf{u},\widehat{\theta}^{(s)}\right)=\frac{\exp\left(\widehat{\theta}_l^{(s)}+\widehat{\theta}_g^{(s)}g_{i_k}\right)}{\sum_{j\in V_t}\exp\left(\widehat{\theta}_l^{(s)}+\widehat{\theta}_g^{(s)}g_j\right)}. \tag{17}$$

For the second term, we obtain

$$p\left(\{i_1,\cdots,i_K\}\setminus\{i_k\}\, are coauthors of q_{c,t} \mid b_{c,t}=i_k,\mathbf{g},\mathbf{u},\widehat{\theta}^{(\Omega)}\right)$$

$$=\prod_{j\in\{i_1,\cdots,i_K\}\setminus i_k} S\left(\widehat{\theta}_c^{(s)}+\gamma\widehat{\theta}_g^{(s)}g_j+\widehat{\theta}_u^{(s)}u_{j,t}\right)\prod_{j\in V_t\setminus\{i_1,\cdots,i_K\}}\left(1-S\left(\widehat{\theta}_c^{(s)}+\gamma\widehat{\theta}_g^{(s)}g_j+\widehat{\theta}_u^{(s)}u_{j,t}t\right)\right). \tag{18}$$

Clearly, $p\left(b_{c,t}=j \mid \mathbf{g},\mathbf{b}\setminus\{b_{c,t}\},\mathbf{q},\mathbf{u},\widehat{\theta}^{(s)}\right)=0$ if $j\in V_t\setminus\{i_1,\cdots,i_K\}$ since a leading author is one of the authors of the paper. By discarding terms unrelated to the authors $\{i_1,\cdots,i_K\}$, we get the practical form of the conditional distribution,

$$p\left(b_{c,t}=i_k \mid \mathbf{g},\mathbf{b}\setminus\{b_{c,t}\},\mathbf{q},\mathbf{u},\widehat{\theta}^{(s)}\right)\propto\frac{\exp\left(\widehat{\theta}_l^{(s)}+\widehat{\theta}_g^{(s)}g_{i_k}\right)}{S\left(\widehat{\theta}_c^{(s)}+\gamma\widehat{\theta}_g^{(s)}g_{i_k}+\widehat{\theta}_u^{(s)}u_{i_k,t}\right)}, k=1,\cdots,K. \tag{19}$$

Then we can choose i_k with probability

$$p\left(b_{c,t}=i_k \mid \mathbf{g},\mathbf{b}\setminus\{b_{c,t}\},\mathbf{q},\mathbf{u},\widehat{\theta}^{(s)}\right)=\frac{\frac{\exp\left(\widehat{\theta}_l^{(s)}+\widehat{\theta}_g^{(s)}g_{i_k}\right)}{S\left(\widehat{\theta}_c^{(s)}+\gamma\widehat{\theta}_g^{(s)}g_{i_k}+\widehat{\theta}_u^{(s)}u_{i_k,t}\right)}}{\sum_{k'=1}^{K}\frac{\exp\left(\widehat{\theta}_l^{(s)}+\widehat{\theta}_g^{(s)}g_{i_{k'}}\right)}{S\left(\widehat{\theta}_c^{(s)}+\gamma\widehat{\theta}_g^{(s)}g_{i_{k'}}+\widehat{\theta}_u^{(s)}u_{i_{k'},t}\right)}} \tag{20}$$

for $k=1,\cdots,K$.

We have the following Gibbs sampling algorithm.

-Initialize: $g_i^{(s,0)}=0, i\in V. b_{c,t}^{(s,0)}$ is assigned randomly from the authors of paper $q_{c,t}, t=1,\cdots,T,\quad c=1,\cdots,c_t$.

-For $h=1,\cdots,H$:

-For $i\in V$:

-Sample $g_i^{(s,h)}$ from $p\left(g_i \mid \mathbf{g}\setminus\{g_i\},\mathbf{b},\mathbf{q},\mathbf{u},\widehat{\theta}^{(s)}\right)$ employing $\mathbf{b}=\mathbf{b}^{(s,h-1)}$.

-For $t=1,\cdots,T$:

-For $c=1,\cdots,c_t$:

doi:10.6342/NTU202501640

-Sample $b_{c,t}^{(s,h)}$ from $p\left(b_{c,t}=i_k \mid \mathbf{g},\mathbf{b}\backslash\{b_{c,t}\},\mathbf{q},\mathbf{u},\widehat{\theta}^{(s)}\right)$ employing $\mathbf{g}=\mathbf{g}^{(s,h)}$ and $b_{c',t'}=b_{c',t'}^{(s,h)}$

if (c',t') is visited and $b_{c',t'}=b_{c',t'}^{(s,h-1)}$ otherwise.

It provides the Gibbs samples $\mathbf{g}^{(s,h)}$ and $\mathbf{b}^{(s,h)}, h=1,\cdots,H$ of genius levels and the leading author assignments, corresponding to the current parameter estimate $\widehat{\theta}^{(s)}$. We ignore some number of samples at the beginning since the initial samples can affect the Gibbs samples. In this paper, we discard the first $H_0=10$ samples out of $H=60$ samples. In addition, the adaptive rejection sampling [8] is used to get samples from the target distribution $p\left(g_i \mid \mathbf{g}\backslash\{g_i\},\mathbf{b},\mathbf{q},\mathbf{u},\widehat{\theta}^{(s)}\right)$ of genius levels.

4.4 EM Algorithm

The E-step estimates the function $Q\left(\theta \mid \widehat{\theta}^{(s)}\right)$ of θ by

$$\widehat{Q}\left(\theta \mid \widehat{\theta}^{(s)}\right)=\frac{1}{H-H_0}\sum_{h=H_0+1}^{H} l\left(\theta \mid \mathbf{g}^{(s,h)},\mathbf{b}^{(s,h)},\mathbf{q},\mathbf{u}\right), \tag{21}$$

The values of $\mathbf{g}^{(s,h)}$ and $\mathbf{b}^{(s,h)}$ are acquired through the Gibbs sampling algorithm. The estimated parameters $\widehat{\theta}^{(s+1)}$ in the M-step are obtained by maximizing the function $\widehat{Q}\left(\theta \mid \widehat{\theta}^{(s)}\right)$ with respect to θ.

$$\widehat{\theta}^{(s+1)}=argmax_{\theta}\widehat{Q}\left(\theta \mid \widehat{\theta}^{(s)}\right). \tag{22}$$

We employ the method of constrained optimization by linear approximation as introduced by (Bos,2006). Ultimately, we arrive at the EM algorithm that iteratively performs E- and M-steps until convergence.

-Initialize: $s=0$.

-For $s=1,2,\cdots$:

-Obtain Gibbs samples $\mathbf{g}^{(s,h)}$ and $\mathbf{b}^{(s,h)}, h=1,\cdots,H$ by running the Gibbs sampling algorithm with $\widehat{\theta}^{(s)}$.

-Calculate $\widehat{Q}\left(\theta \mid \widehat{\theta}^{(s)}\right)$.

-Find $\widehat{\theta}^{(s+1)}=argmax_{\theta}\widehat{Q}\left(\theta \mid \widehat{\theta}^{(s)}\right)$.

-If converged:

- $\widehat{\theta}=\widehat{\theta}^{(s+1)}$.

-Obtain Gibbs samples $\mathbf{g}^{(h)}$ and $\mathbf{b}^{(h)}, h=1,\cdots,H$ by running the Gibbs sampling algorithm with

$\widehat{\theta}$.

-Break

The algorithm yields the converged parameter estimate $\widehat{\theta}$ and the corresponding Gibbs samples $\mathbf{g}^{(h)}$ and $\mathbf{b}^{(h)}, h=1,\cdots,H$.

doi:10.6342/NTU202501640

4.5 Inference

According to Louis' method (Louis, 1982), the estimated observed information matrix of the estimated model parameter $\hat{\theta}$ can be expressed as follows:

$$\widehat{I}(\widehat{\theta}) = -\nabla^2 Q(\widehat{\theta} \mid \widehat{\theta}) + [\nabla Q(\widehat{\theta} \mid \widehat{\theta})][\nabla Q(\widehat{\theta} \mid \widehat{\theta})]' - \frac{1}{H - H_0} \sum_{h=H_0+1}^{H} \left[l\left(\widehat{\theta} \mid \mathbf{g}^{(h)}, \mathbf{b}^{(h)}, \mathbf{q}, \mathbf{u}\right) \right] \left[l\left(\widehat{\theta} \mid \mathbf{g}^{(h)}, \mathbf{b}^{(h)}, \mathbf{q}, \mathbf{u}\right) \right]', \quad (23$$

By using the differential operator $\nabla = \left(\frac{\partial}{\partial \theta_l}, \frac{\partial}{\partial \theta_c}, \frac{\partial}{\partial \theta_g}, \frac{\partial}{\partial \theta_u} \right)$ on the parameter vector, we can compute the standard error vector of the parameter estimate by

$$s.e.(\widehat{\theta}) = \left(\sqrt{\left[\widehat{I}(\widehat{\theta})^{-1} \right]_{\mathrm{xx}}} \right)_{x=1,2,3,4}. \quad (24)$$

Using the estimated posterior distributions, we can estimate the genius levels of the authors.

$$\widehat{g}_i = \frac{1}{H - H_0} \sum_{h=H_0+1}^{H} g_i^{(h)}, \quad i \in V \quad (25)$$

Likewise, we can use the estimated posterior distributions of $b_{c,t}$ to select the most probable leading author of a paper $q_{c,t}$ by determining the most commonly occurring author among the posterior Gibbs samples $\left(b_{c,t}^{(h)} \right)_{h=H|_0+1,\cdots,H}$.We may express it by

$$\widehat{b}_{c,t} = argmax_{i \in V_t} \frac{1}{H - H_0} \sum_{h=H_0+1}^{H} \mathbf{1}\left(b_{c,t}^{(h)} = i \right), \quad t = 1, \cdots, T, c = 1, \cdots, c_t \quad (26)$$

The 1(statement) function is an indicator f unction that returns a value of 1 if the statement is true and 0 if the statement is false.

5 Real Data Analyze

The leading author model [14] focused on the two subjects, Management and Statistics Probability, from 2007 to 2016. In this study, we extend the previous work by considering more subjects and time periods to achieve consistency and completeness. We now consider 183 subjects and 35 years (1981-2016). Likewise, in [14], we evaluate the genius levels of the authors with publications at more than or equal to two years. The genius levels are set to 0 for the authors who appear at only one year.

For a comparison study among various subjects, we first consider the four parameters of the model:

1. The parameter $\hat{\theta}_l$ controls the total number of papers per unit time.

doi:10.6342/NTU202501640

2. The genius coefficient $\hat{\theta_g}$ determines the effect size of authors' genius levels. High $\hat{\theta_g}$ indicates that the author heterogeneity is relatively large in the subject.
3. The parameter $\hat{\theta_c}$ controls the number of authors per paper.
4. The popularity coefficient $\hat{\theta_u}$ determines the magnitude of the popularity effect. A positive $\hat{\theta_u}$ indicates that the rich-get-richer effect is working on the collaboration system, whereas a negative $\hat{\theta_u}$ indicates the opposite direction, rich-get-poorer. The parameter $\hat{\theta_u}$ of the subject with a high popularity effect will have a large positive value.

1.Parameter Estimates We apply the model to 182 subjects and obtain the estimates of the model parameters and the genius levels. Table 1 shows the estimates of the model parameters.

Table 1: Parameter estimates for 182 subjects

Subjects	$\hat{\theta_l}$	$\hat{\theta_g}$	$\hat{\theta_c}$	$\hat{\theta_u}$	s.e.(θ_l)	s.e.(θ_g)	s.e.(θ_c)	s.e.(θ_u)
Acoustics	-1.3798	0.2963	-8.9042	0.2306	0.0065	0.0041	0.0042	0.0041
Agricultural Economics & Policy	-1.3599	0.2896	-9.0801	0.2243	0.0060	0.0038	0.0040	0.0040
Agricultural Engineering	-1.5407	0.3060	-8.5270	0.2176	0.0073	0.0043	0.0041	0.0064
Agriculture, Dairy & Animal Science	-1.6144	0.3662	-9.6153	0.1674	0.0043	0.0021	0.0023	0.0011
Agriculture, Multidisciplinary	-1.6815	0.3220	-9.2014	0.0768	0.0054	0.0031	0.0030	0.0053
Agriculture	-1.6157	0.3730	-10.7356	0.2100	0.0026	0.0012	0.0014	0.0007
Agronomy	-1.6797	0.3010	-9.4584	0.2294	0.0048	0.0027	0.0027	0.0035
Allergy	-1.6279	0.4781	-8.9444	0.1886	0.0057	0.0024	0.0029	0.0010
Anatomy & Morphology	-1.6476	0.3502	-8.0012	0.0195	0.0097	0.0063	0.0051	0.0041
Andrology	-1.6830	0.3562	-8.2506	-0.0062	0.0086	0.0053	0.0045	0.0041
Anesthesiology	-1.5096	0.5065	-9.4786	0.0317	0.0044	0.0026	0.0025	0.0020
Anthropology	-1.2987	0.4772	-10.0521	0.0864	0.0035	0.0022	0.0023	0.0018
Archaeology	-1.0412	0.3447	-8.6322	-0.3492	0.0083	0.0067	0.0081	0.0209
Architecture	-0.9148	0.3374	-9.0970	-0.2204	0.0068	0.0058	0.0071	0.0175
Area Studies	-0.6682	0.2624	-10.1237	-0.1541	0.0047	0.0047	0.0067	0.0177
Arts & Humanities - Other Topics	-0.4706	0.3541	-11.0333	-0.5676	0.0034	0.0034	0.0062	0.0182
Art	-0.5146	0.3413	-11.1424	-0.7119	0.0033	0.0032	0.0059	0.0186
Asian Studies	-0.2708	0.2964	-10.0495	-2.1627	0.0097	0.0106	0.0381	0.2099
Astronomy & Astrophysics	-2.2722	0.8865	-9.5281	0.0135	0.0034	0.0007	0.0013	0.0001
Audiology & Speech-Language Pathology	-2.2979	0.8982	-9.7696	0.0165	0.0032	0.0007	0.0013	0.0001
Automation & Control Systems	-1.1721	0.4006	-9.5215	0.2430	0.0050	0.0030	0.0036	0.0019
Behavioral Sciences	-1.5443	0.3153	-9.3391	0.1529	0.0051	0.0032	0.0031	0.0042
Biochemical Research Methods	-1.6542	0.3694	-10.0775	0.2087	0.0034	0.0018	0.0018	0.0013
Biodiversity Conservation	-1.6208	0.2818	-8.9499	0.1285	0.0063	0.0039	0.0037	0.0063
Biology	-1.5205	0.3860	-10.1004	0.0558	0.0033	0.0022	0.0020	0.0032
Biomedical Social Sciences	-1.3057	0.3987	-8.7984	0.0305	0.0067	0.0047	0.0046	0.0086
Biophysics	-1.6201	0.3733	-10.4070	0.2691	0.0029	0.0014	0.0016	0.0008
Biotechnology & Applied Microbiology	-1.7278	0.3956	-10.6876	0.2217	0.0025	0.0011	0.0013	0.0005
Business & Economics	-0.8488	0.3661	-11.2020	0.2798	0.0023	0.0016	0.0022	0.0014
Business, Finance	-1.0063	0.1548	-9.1884	0.3912	0.0070	0.0052	0.0074	0.0112
Business	-1.0084	0.2570	-10.0632	0.3400	0.0044	0.0032	0.0042	0.0046
Cardiac & Cardiovascular Systems	-1.5278	0.8133	-10.3497	0.0091	0.0022	0.0009	0.0009	0.0001

doi:10.6342/NTU202501640

Cardiovascular System & Cardiology	-1.5504	0.9329	-10.6366	-0.0135	0.0019	0.0008	0.0008	0.0001
Cell & Tissue Engineering	-1.8825	0.3824	-8.5348	0.0899	0.0070	0.0035	0.0032	0.0038
Cell Biology	-1.8348	0.3815	-10.9555	0.2312	0.0022	0.0009	0.0010	0.0004
Chemistry, Analytical	-1.5229	0.4689	-10.3166	0.1596	0.0030	0.0014	0.0016	0.0006
Chemistry, Applied	-1.5853	0.4008	-9.9276	0.2125	0.0037	0.0021	0.0021	0.0016
Chemistry, Inorganic & Nuclear	-1.4953	0.4053	-9.7992	0.2212	0.0039	0.0017	0.0022	0.0007
Chemistry, Medicinal	-1.6643	0.4275	-10.4981	0.1956	0.0027	0.0011	0.0014	0.0004
Chemistry, Multidisciplinary	-1.3972	0.6910	-11.2465	0.0570	0.0017	0.0008	0.0009	0.0002
Chemistry, Organic	-1.4125	0.9235	-11.4371	0.0177	0.0015	0.0008	0.0008	0.0002
Chemistry, Physical	-1.3943	0.5710	-11.1181	0.1305	0.0019	0.0008	0.0010	0.0002
Classics	-0.2574	0.1822	-10.6875	0.2724	0.0114	0.0131	0.0622	0.1078
Clinical Neurology	-1.5706	0.5922	-10.9088	0.1336	0.0021	0.0008	0.0010	0.0002
Communication	-0.7586	0.3162	-9.0949	0.0347	0.0073	0.0068	0.0087	0.0165
Computer Science, Artificial Intelligence	-1.1466	0.3891	-9.9058	0.2624	0.0040	0.0024	0.0029	0.0016
Computer Science, Cybernetics	-1.1567	0.4239	-9.9698	0.2203	0.0039	0.0023	0.0028	0.0017
Computer Science, Hardware & Architecture	-1.3150	0.4030	-8.9950	0.0799	0.0062	0.0039	0.0043	0.0059
Computer Science, Information Systems	-1.2069	0.4070	-10.2075	0.2531	0.0034	0.0020	0.0024	0.0016
Computer Science, Interdisciplinary Applications	-1.2305	0.4756	-10.7303	0.1630	0.0026	0.0016	0.0018	0.0012
Computer Science, Software Engineering	-1.2834	0.3378	-9.5632	0.2532	0.0049	0.0031	0.0035	0.0033
Computer Science, Theory & Methods	-1.2216	0.3468	-9.4990	0.2314	0.0050	0.0034	0.0038	0.0051
Computer Science	-1.2140	0.4172	-11.2650	0.2447	0.0021	0.0010	0.0015	0.0005
Construction & Building Technology	-1.1628	0.3686	-9.3284	0.2408	0.0052	0.0034	0.0038	0.0039
Criminology & Penology	-1.1324	0.3782	-9.6538	0.2273	0.0046	0.0030	0.0034	0.0032
Critical Care Medicine	-1.6241	0.4747	-9.8890	0.2060	0.0036	0.0016	0.0019	0.0008
Crystallography	-1.6227	0.4470	-9.1131	0.0367	0.0056	0.0025	0.0030	0.0004
Cultural Studies	-0.3518	0.0790	-9.0361	-1.0171	0.0114	0.0162	0.0262	0.1849
Dance	-0.3630	0.2163	-9.1362	0.3726	0.0108	0.0151	0.0212	0.0239
Demography	-0.6164	0.0521	-9.0842	0.3449	0.0085	0.0095	0.0116	0.0188
Dentistry, Oral Surgery & Medicine	-1.5550	0.4067	-9.4991	0.2141	0.0046	0.0022	0.0026	0.0013
Dermatology	-1.4984	0.4706	-10.4109	0.1656	0.0028	0.0013	0.0016	0.0005
Developmental Biology	-1.7675	0.3507	-8.8520	0.0356	0.0063	0.0036	0.0035	0.0056
Ecology	-1.4815	0.3179	-10.2437	0.3113	0.0033	0.0017	0.0020	0.0014
Economics	-0.8501	0.4073	-10.6566	0.1997	0.0030	0.0021	0.0029	0.0018
Education & Educational Research	-0.8940	0.3828	-11.1507	0.1322	0.0022	0.0016	0.0020	0.0016
Education, Scientific Disciplines	-1.3046	0.3931	-8.7819	-0.0532	0.0068	0.0052	0.0047	0.0088
Education, Special	-1.2728	0.4025	-7.9681	0.1645	0.0102	0.0067	0.0067	0.0034
Electrochemistry	-1.4738	0.4179	-9.8431	0.2221	0.0038	0.0017	0.0021	0.0010
Emergency Medicine	-1.5071	0.4323	-10.1151	0.2226	0.0033	0.0016	0.0018	0.0009
Endocrinology & Metabolism	-1.7539	0.4689	-10.5109	0.1618	0.0027	0.0012	0.0013	0.0005
Energy & Fuels	-1.3199	0.5101	-10.4515	0.1603	0.0027	0.0014	0.0016	0.0009
Engineering, Aerospace	-1.2858	0.3382	-8.4252	0.0462	0.0083	0.0057	0.0059	0.0116
Engineering, Biomedical	-1.5720	0.3216	-9.9323	0.2048	0.0036	0.0019	0.0020	0.0018
Engineering, Chemical	-1.3658	0.4375	-10.6123	0.2376	0.0026	0.0013	0.0016	0.0007
Engineering, Civil	-1.1916	0.4246	-10.1509	0.1491	0.0035	0.0023	0.0025	0.0024
Engineering, Electrical & Electronic	-1.2601	0.5398	-11.1263	0.1208	0.0020	0.0010	0.0013	0.0004
Engineering, Environmental	-1.4927	0.3344	-9.8014	0.2935	0.0040	0.0022	0.0023	0.0019

doi:10.6342/NTU202501640

Engineering, Geological	-1.4237	0.3864	-10.0358	0.2671	0.0035	0.0021	0.0022	0.0018
Engineering, Industrial	-1.2226	0.3421	-8.9749	0.1002	0.0065	0.0046	0.0048	0.0077
Engineering, Manufacturing	-1.2170	0.3509	-9.5598	0.2187	0.0047	0.0031	0.0034	0.0043
Engineering, Marine	-1.2339	0.3937	-9.6481	0.1544	0.0045	0.0031	0.0033	0.0043
Engineering, Mechanical	-1.1857	0.3826	-10.2538	0.2446	0.0033	0.0020	0.0024	0.0019
Engineering, Multidisciplinary	-1.1098	0.3785	-10.7524	0.2737	0.0025	0.0016	0.0019	0.0012
Engineering, Ocean	-1.3647	0.3273	-7.6443	-0.0564	0.0125	0.0082	0.0086	0.0200
Engineering, Petroleum	-1.3869	0.3024	-7.7322	0.1073	0.0112	0.0080	0.0071	0.0131
Entomology	-1.4012	0.3537	-9.3474	0.1640	0.0051	0.0032	0.0033	0.0044
Environmental Sciences & Ecology	-1.4928	0.3560	-11.4593	0.3123	0.0018	0.0008	0.0011	0.0004
Environmental Sciences	-1.5402	0.3705	-11.0711	0.3173	0.0022	0.0010	0.0013	0.0005
Environmental Studies	-1.1199	0.3134	-9.5088	-0.0752	0.0050	0.0040	0.0042	0.0103
Ergonomics	-1.3171	0.3215	-7.7306	0.0947	0.0118	0.0089	0.0082	0.0154
Ethics	-0.7083	0.3817	-8.8909	0.1056	0.0083	0.0079	0.0099	0.0143
Ethnic Studies	-0.6815	0.1954	-7.9563	-0.2316	0.0144	0.0172	0.0197	0.0731
Evolutionary Biology	-1.4543	0.3040	-9.4975	0.2539	0.0049	0.0030	0.0033	0.0036
Family Studies	-1.3925	0.2793	-9.8122	0.2897	0.0042	0.0027	0.0029	0.0034
Film, Radio & Television	-1.3421	0.2646	-9.9250	0.3021	0.0041	0.0026	0.0029	0.0031
Film, Radio, Television	-1.3451	0.2946	-9.9177	0.2552	0.0041	0.0026	0.0029	0.0032
Fisheries	-1.5877	0.3576	-8.9969	0.2405	0.0061	0.0032	0.0036	0.0036
Folklore	-1.5243	0.3405	-9.1002	0.2490	0.0058	0.0032	0.0036	0.0036
Food Science & Technology	-1.5414	0.4638	-10.2846	0.1909	0.0031	0.0014	0.0017	0.0007
Forestry	-1.5153	0.2572	-8.9854	0.2833	0.0063	0.0038	0.0040	0.0046
Gastroenterology & Hepatology	-1.6649	0.6483	-10.3993	0.0968	0.0027	0.0011	0.0013	0.0003
General & Internal Medicine	-1.5950	0.4048	-11.2466	0.2751	0.0019	0.0008	0.0010	0.0005
Genetics & Heredity	-1.9456	0.4093	-10.3991	0.2101	0.0028	0.0011	0.0013	0.0005
Geochemistry & Geophysics	-1.4879	0.3332	-9.6993	0.2667	0.0044	0.0023	0.0028	0.0020
Geosciences, Multidisciplinary	-1.5044	0.3705	-10.3761	0.2062	0.0030	0.0015	0.0018	0.0013
Health Care Sciences & Services	-1.4071	0.4834	-10.0412	0.1098	0.0033	0.0019	0.0019	0.0014
Health Policy & Services	-1.4087	0.4825	-10.0858	0.1152	0.0033	0.0020	0.0020	0.0014
Hematology	-1.7106	0.6919	-10.8767	0.1008	0.0021	0.0009	0.0010	0.0002
History & Philosophy Of Science	-0.4583	0.3211	-9.4269	-1.4768	0.0075	0.0081	0.0142	0.0569
History	-0.2486	0.3098	-11.8073	-1.9766	0.0035	0.0036	0.0129	0.0499
Hospitality, Leisure, Sport & Tourism	-0.3632	0.3795	-11.4152	0.0957	0.0032	0.0031	0.0061	0.0051
Humanities, Multidisciplinary	-0.3384	0.3753	-11.8234	-0.0225	0.0028	0.0026	0.0059	0.0065
Information Science & Library Science	-0.7982	0.7641	-9.5299	-0.3948	0.0057	0.0022	0.0058	0.0067
Instruments & Instrumentation	-1.5969	0.3785	-9.8914	0.2638	0.0034	0.0016	0.0019	0.0011
Law	-0.5563	0.2869	-9.6800	0.0199	0.0065	0.0066	0.0103	0.0202
Legal Medicine	-0.9493	0.3879	-9.4602	0.1566	0.0055	0.0041	0.0048	0.0029
Literature, British Isles	-0.8968	0.3906	-9.5752	0.1725	0.0052	0.0039	0.0047	0.0029
Literature, German, Dutch, Scandinavian	-0.8591	0.3834	-9.7091	0.1397	0.0050	0.0038	0.0047	0.0031
Literature, Romance	-0.7463	0.4380	-10.1447	0.0743	0.0044	0.0036	0.0047	0.0031
Management	-1.0529	0.2655	-9.8867	0.2837	0.0046	0.0034	0.0043	0.0051
Materials Science, Ceramics	-1.1997	0.3844	-10.1090	0.2428	0.0036	0.0024	0.0027	0.0024
Materials Science, Characterization & Testing	-1.2194	0.3522	-10.2881	0.2530	0.0033	0.0021	0.0025	0.0023
Materials Science, Composites	-1.2255	0.3439	-10.4638	0.2981	0.0031	0.0019	0.0022	0.0018

doi:10.6342/NTU202501640

Materials Science	-1.2753	0.7915	-11.2195	-0.0071	0.0014	0.0006	0.0007	0.0001
Mathematical & Computational Biology	-1.4889	0.2518	-9.2610	0.2277	0.0055	0.0035	0.0035	0.0052
Mathematics, Applied	-0.8847	0.4765	-10.8095	0.1230	0.0029	0.0018	0.0027	0.0010
Mechanics	-1.1183	0.3960	-10.3024	0.2325	0.0033	0.0020	0.0025	0.0017
Mineralogy	-1.1867	0.4108	-10.4311	0.2182	0.0031	0.0019	0.0022	0.0016
Mycology	-1.3068	0.4086	-10.4893	0.2210	0.0030	0.0017	0.0020	0.0014
Neurosciences & Neurology	-1.6610	0.7520	-11.4594	0.0001	0.0016	0.0007	0.0008	0.0002
Neurosciences	-1.6267	0.4972	-11.1592	0.1619	0.0020	0.0008	0.0010	0.0004
Nuclear Science & Technology	-1.6991	0.3728	-9.4180	0.1810	0.0046	0.0022	0.0024	0.0017
Oceanography	-1.5683	0.2732	-9.2609	0.2449	0.0055	0.0032	0.0036	0.0051
Operations Research & Management Science	-1.0735	0.3451	-9.6281	0.2695	0.0049	0.0033	0.0040	0.0031
Pathology	-1.6166	0.5847	-9.8411	0.1278	0.0035	0.0015	0.0017	0.0006
Philosophy	-0.2504	0.3515	-10.8581	-0.6387	0.0054	0.0040	0.0169	0.0414
Physics, Atomic, Molecular & Chemical	-1.3801	0.4016	-10.1209	0.1894	0.0034	0.0018	0.0022	0.0012
Physics, Condensed Matter	-1.5692	0.3857	-10.5556	0.2151	0.0027	0.0012	0.0015	0.0006
Physics, Multidisciplinary	-1.7554	0.7092	-10.2272	0.1565	0.0031	0.0009	0.0016	0.0002
Physics	-1.9098	0.8762	-11.8565	0.0184	0.0013	0.0003	0.0007	0.0000
Physiology	-1.6522	0.3867	-9.8124	0.0996	0.0038	0.0021	0.0021	0.0017
Primary Health Care	-1.3363	0.4541	-8.0758	0.0204	0.0093	0.0065	0.0060	0.0089
Psychiatry	-1.4503	0.5197	-10.5465	0.1605	0.0026	0.0012	0.0015	0.0004
Psychology, Experimental	-1.1388	0.3227	-9.6048	0.2712	0.0046	0.0029	0.0034	0.0020
Psychology, Mathematical	-1.2029	0.3910	-6.9828	-0.2655	0.0183	0.0142	0.0157	0.0351
Psychology, Multidisciplinary	-1.2258	0.3814	-9.7823	0.0833	0.0040	0.0027	0.0028	0.0036
Public, Environmental & Occupational Health	-1.6349	0.5632	-10.6668	0.0918	0.0025	0.0014	0.0013	0.0008
Radiology, Nuclear Medicine & Medical Imaging	-1.6625	0.4991	-10.4072	0.1551	0.0027	0.0011	0.0014	0.0004
Rehabilitation	-1.6113	0.5070	-10.7692	0.1506	0.0023	0.0010	0.0012	0.0004
Religion	-0.2720	0.5169	-10.4766	-1.6649	0.0058	0.0041	0.0157	0.0590
Rheumatology	-1.3298	0.6883	-9.9726	0.0853	0.0034	0.0016	0.0019	0.0004
Robotics	-1.3585	0.2389	-7.9214	0.2188	0.0107	0.0073	0.0073	0.0125
Science & Technology - Other Topics	-1.8087	0.5207	-11.5542	0.0309	0.0015	0.0005	0.0007	0.0002
Social Issues	-0.7669	0.3741	-8.5800	0.0613	0.0094	0.0096	0.0105	0.0177
Social Sciences - Other Topics	-0.8251	0.3401	-10.2572	0.1604	0.0038	0.0033	0.0038	0.0046
Social Sciences, Biomedical	-1.3079	0.3863	-8.7914	0.0410	0.0067	0.0046	0.0046	0.0087
Social Sciences, Interdisciplinary	-1.1188	0.3068	-9.7551	0.1773	0.0044	0.0034	0.0035	0.0065
Social Sciences, Mathematical Methods	-1.0745	0.3174	-10.0222	0.1543	0.0039	0.0031	0.0033	0.0057
Social Work	-1.0778	0.3079	-8.3915	-0.0441	0.0090	0.0073	0.0078	0.0157
Sociology	-0.7933	0.2672	-10.0634	-0.1155	0.0045	0.0042	0.0056	0.0125
Soil Science	-1.0942	0.2752	-10.2205	0.2439	0.0037	0.0029	0.0032	0.0040
Spectroscopy	-1.3107	0.3834	-10.4851	0.1939	0.0031	0.0019	0.0021	0.0016
Sport Sciences	-1.4495	0.4257	-9.6858	0.2232	0.0042	0.0022	0.0025	0.0012
Statistics & Probability	-1.0433	0.3800	-9.7697	0.1894	0.0046	0.0029	0.0038	0.0016
Substance Abuse	-1.1886	0.3941	-10.1003	0.2335	0.0037	0.0020	0.0026	0.0011
Surgery	-1.5544	0.6320	-11.2933	0.0458	0.0018	0.0008	0.0009	0.0003
Telecommunications	-1.1775	0.4482	-9.9324	0.2229	0.0038	0.0019	0.0027	0.0010
Theater	-1.1318	0.4592	-10.0412	0.2244	0.0037	0.0019	0.0026	0.0010
Thermodynamics	-1.1766	0.4469	-10.6316	0.2376	0.0028	0.0014	0.0019	0.0007

doi:10.6342/NTU202501640

Toxicology	-1.7167	0.4015	-9.8035	0.1888	0.0038	0.0017	0.0019	0.0010
Transplantation	-1.7767	0.6519	-9.6444	0.0807	0.0037	0.0014	0.0017	0.0005
Transportation Science & Technology	-1.1965	0.3236	-8.7297	0.3059	0.0073	0.0046	0.0053	0.0043
Transportation	-1.1258	0.3828	-9.0667	0.2616	0.0061	0.0039	0.0045	0.0033
Tropical Medicine	-1.5248	0.3689	-9.4895	0.1977	0.0046	0.0026	0.0027	0.0022
Urban Studies	-1.4057	0.4132	-9.7614	0.2058	0.0041	0.0026	0.0026	0.0022
Urology & Nephrology	-1.5961	0.6594	-10.1359	0.0881	0.0031	0.0012	0.0017	0.0002
Veterinary Sciences	-1.5771	0.5862	-10.6465	0.0946	0.0024	0.0009	0.0012	0.0002
Virology	-1.6409	0.5858	-10.8895	0.0954	0.0021	0.0008	0.0011	0.0002
Water Resources	-1.4106	0.3839	-9.9751	0.2288	0.0038	0.0023	0.0024	0.0023
Women's Studies	-1.3143	0.3561	-10.1149	0.2442	0.0034	0.0021	0.0023	0.0023
Zoology	-1.3243	0.4086	-10.0633	0.1660	0.0036	0.0022	0.0024	0.0020

2.Log t- rank plot

For each parameter, we perform the t-tests for all the possible 16,471 pairs of 182 subjects. Figure 1 shows the log t-rank plot using the logarithm of t-statistic with base 2 for each subject pair. We apply the K-Means clustering algorithm to the log-t-statistics values and depict it with red lines. We can see that the curves are almost the same, indicating the four parameter estimates have positive relationships.

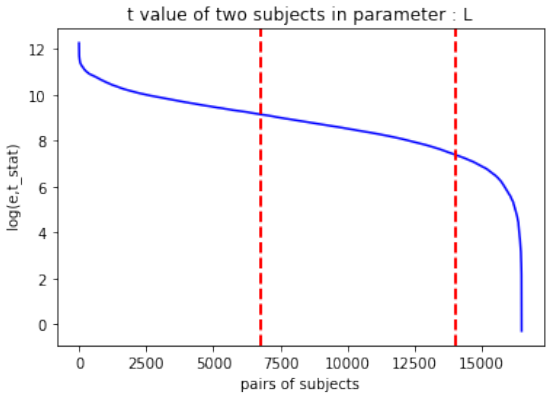


Figure 1: Parameter $\hat{\theta}_l$ (t distribution order graph).

doi:10.6342/NTU202501640

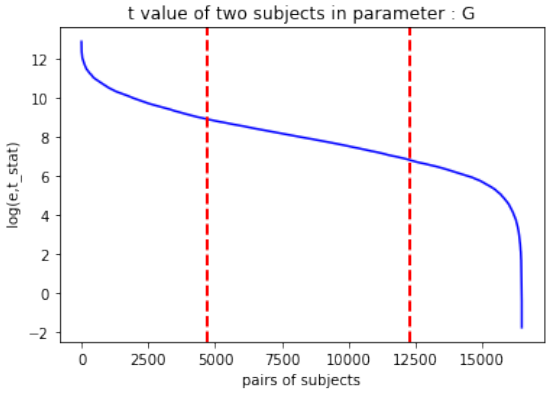


Figure 2: Parameter $\hat{\theta_g}$ (t distribution order graph).

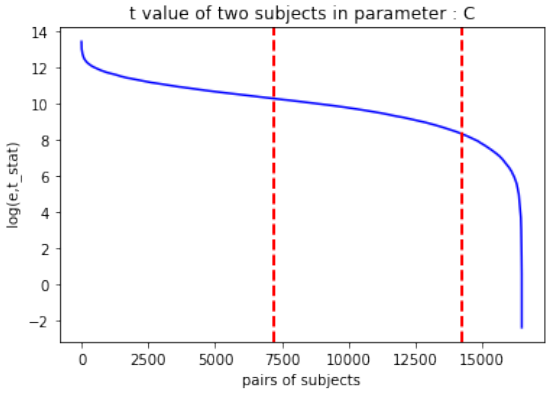


Figure 3: Parameter $\hat{\theta_c}$ (t distribution order graph).

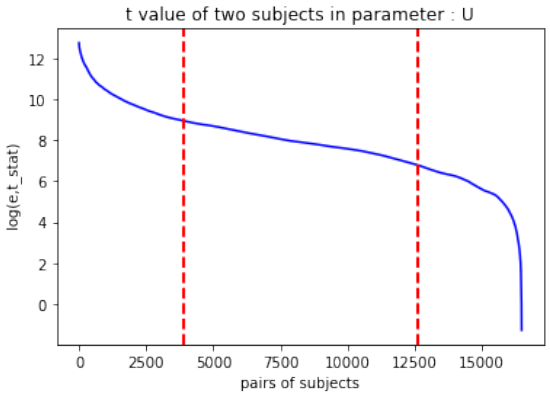


Figure 4: Parameter $\hat{\theta_u}$ (t distribution order graph).

doi:10.6342/NTU202501640

3.Parameter Rank Correlation matrix

Thus, we calculate Spearman's rank correlation to check the correlations between the two coefficients across the subjects. Figure 2 shows the scatter plot matrix of the four model parameters with Spearman's rank correlation values. All four parameters turn out to have positive correlations.

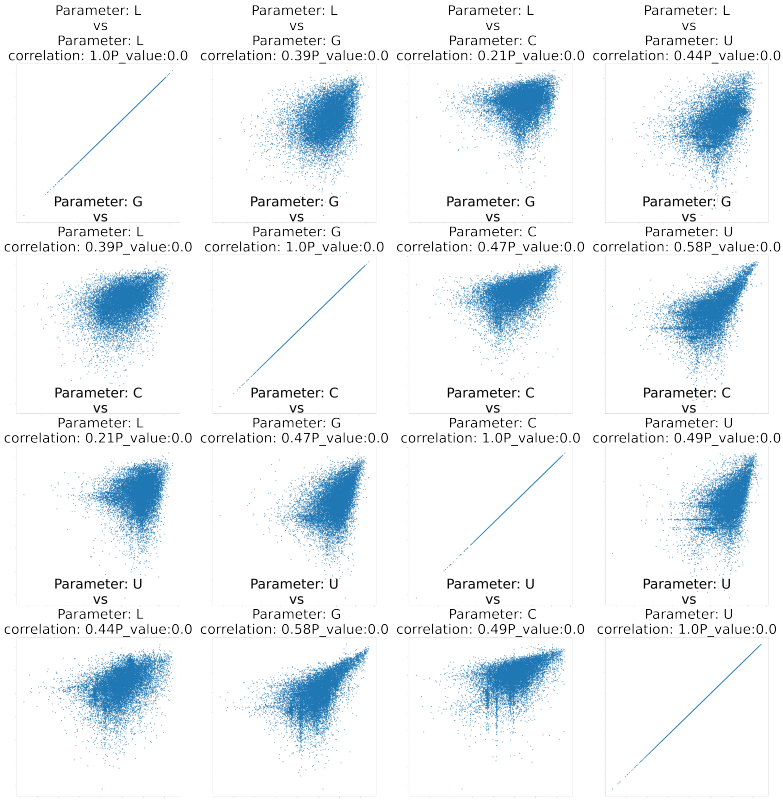


Figure 5: Spearman Rank Correlation

4.Subject Similarity Analysis and SSS plot

We perform a subject similarity analysis based on a specific subject, where the similarity is based on the model parameter estimates. We regard the two subjects are similar if the rich-get-richer nature and paper production mechanism are similar. We place subjects close to the target subject if they are similar. The following steps illustrate how to create a similar subject spiral(SSS) plot for a target subject.

Step 1:
Set the target subject A and non-target subjects B_j , j= 1,2,...,182

Step 2:
For each parameter θ_i, $i = l, g, c, u$ calculate the t-statistics $T_{i,j}$ for a non-target subject B_j

doi:10.6342/NTU202501640

H_0: There is no significant difference in the parameters θi between the groups of data of two subjects.

H_1: There is a significant difference in the parameters θi between the groups of data of two subjects.

In this step, we employ a t-test to examine whether there is a statistical difference between the two subjects in terms of the rich-get-richer nature or the paper production mechanism.

Step 3:
For each parameter θ_i ,calculate the rank $R_{i,j}$ of a non-target subject B_j based on the similarity.

Step 4:
Determine the distance D_j=$\frac{1}{4}$ ($R_{l,j}$+$R_{g,j}$+$R_{c,j}$+$R_{u,j}$) for each non-target subject B_j by averaging the rank values.

Step 5:
Standardize the distances d_j, j=1,2,⋯,182.

Step 6:
Place the target node in the center and arrange the non-target nodes in a spiral shape according to the standardized distances.

Figure 6 shows the SSS plot for the target subject, Business. Interestingly, the Dance subject turns out to be the closest subject to the target subject. Subjects such as Management, Social Issues, Law, Business, Finance, and Communication are also shown to be similar to the target field, which appears to be due to similarities in the backgrounds of the subjects.

doi:10.6342/NTU202501640

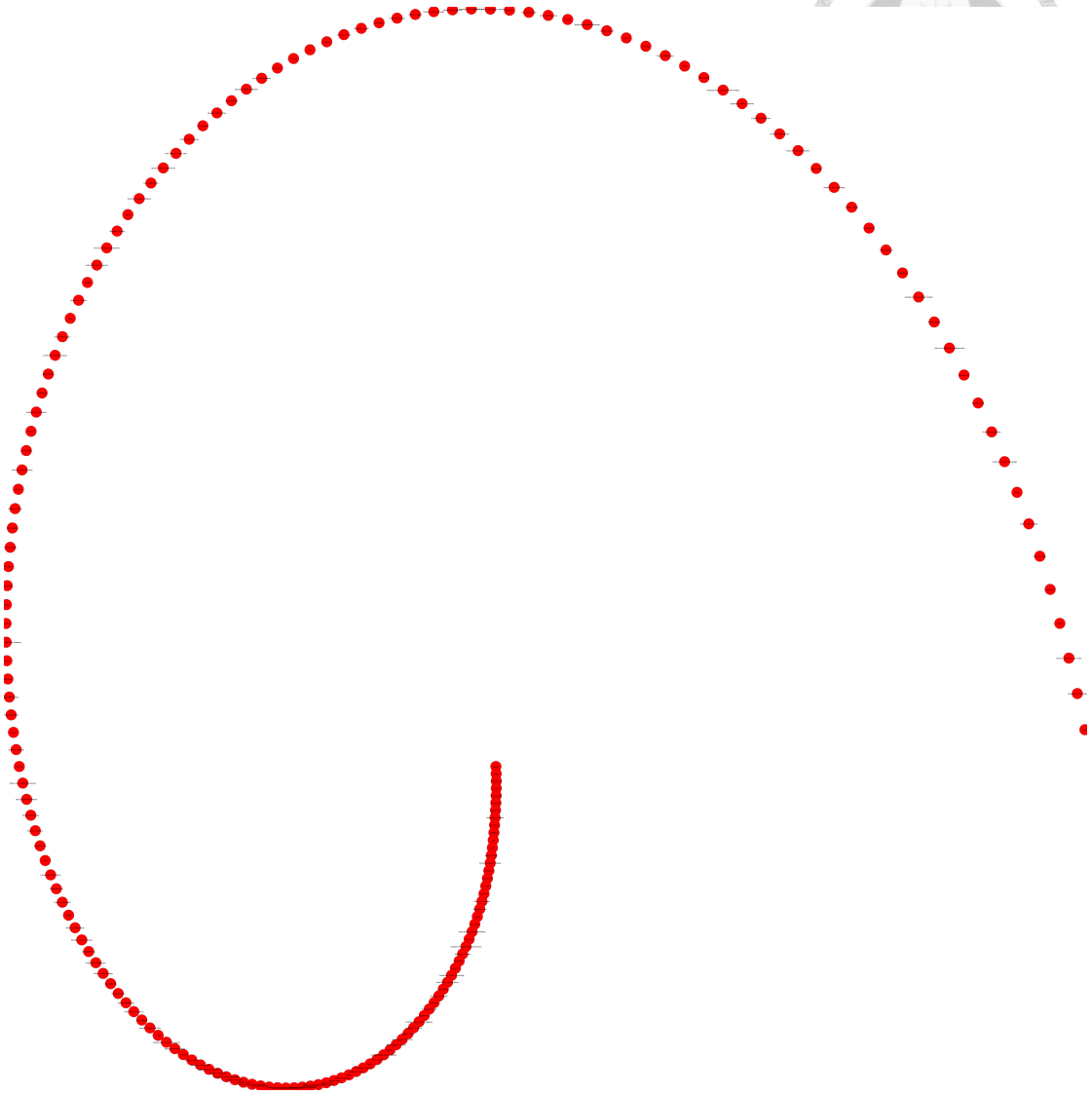


Figure 6: SSS plot for *Business*

doi:10.6342/NTU202501640

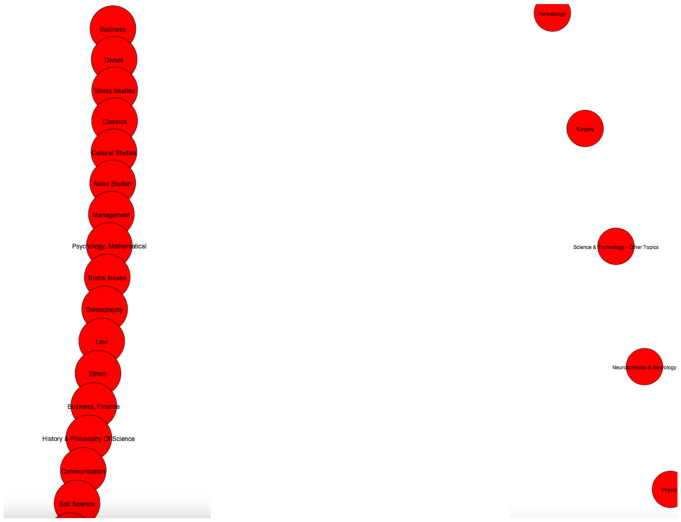


Figure 7: the most similar subjects (left) and the least similar subjects(right) with Business

doi:10.6342/NTU202501640

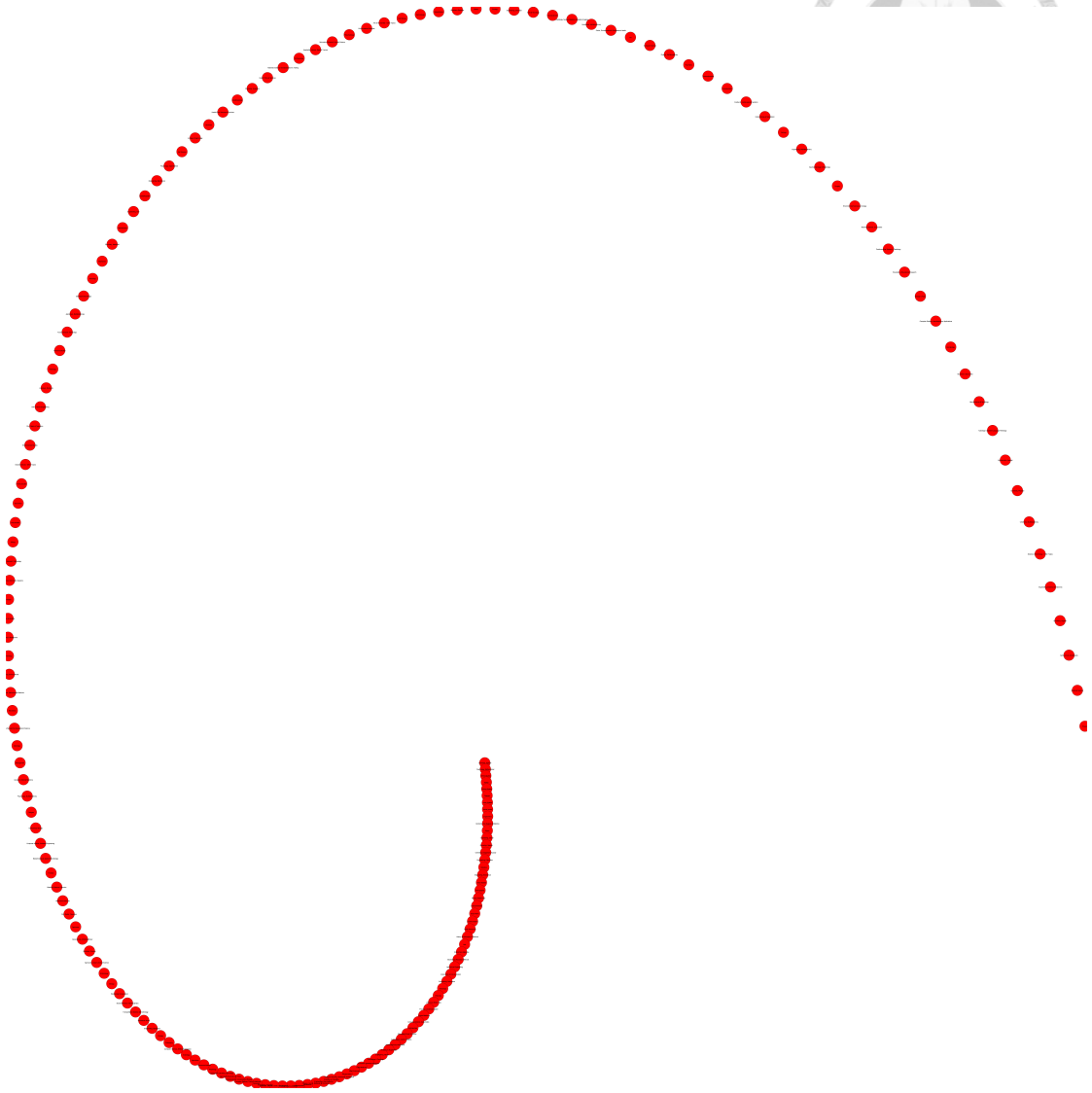


Figure 8: Similar subjects graph for Chemistry, Applied

doi:10.6342/NTU202501640

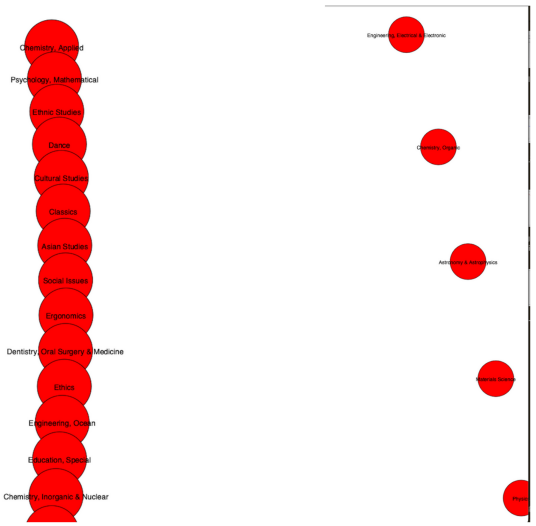


Figure 9: the most similar subjects (left) and the least similar subjects(right) with Chemistry, Applied

doi:10.6342/NTU202501640

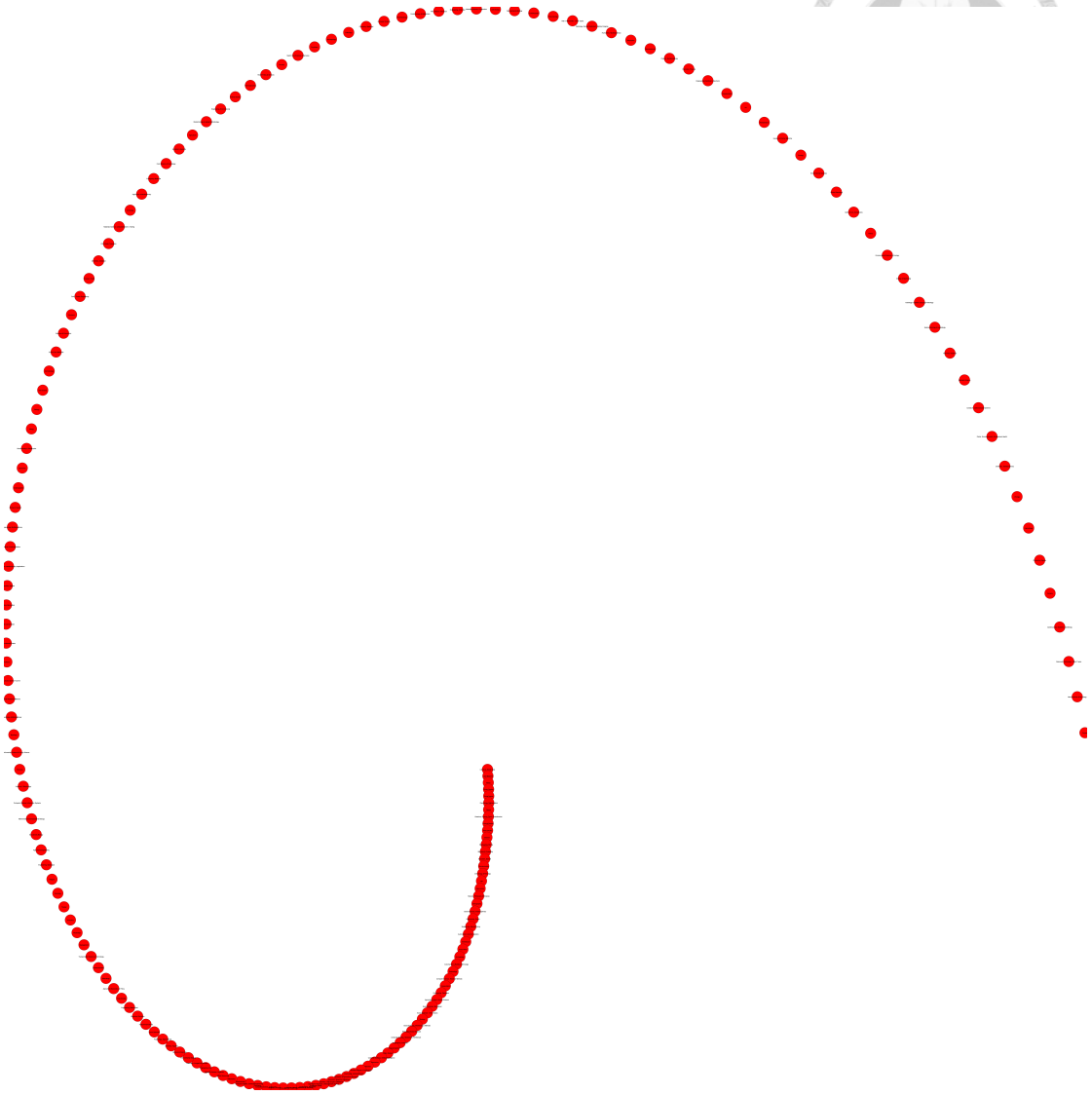


Figure 10: Similar subjects graph for Literature, British Isles

doi:10.6342/NTU202501640

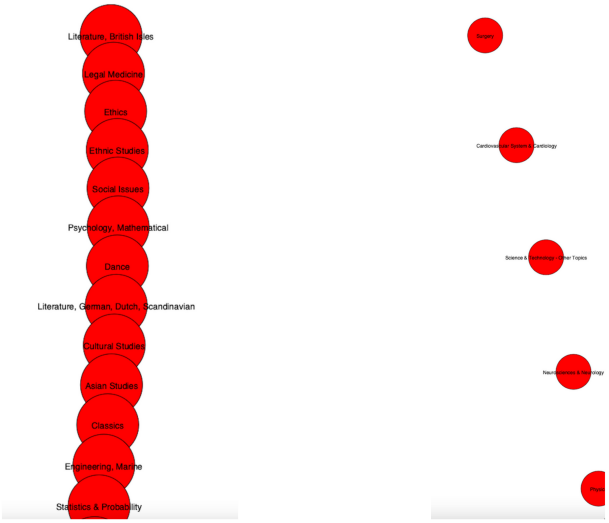


Figure 11: the most similar subjects (left) and the least similar subjects(right) with Literature, British Isles

doi:10.6342/NTU202501640

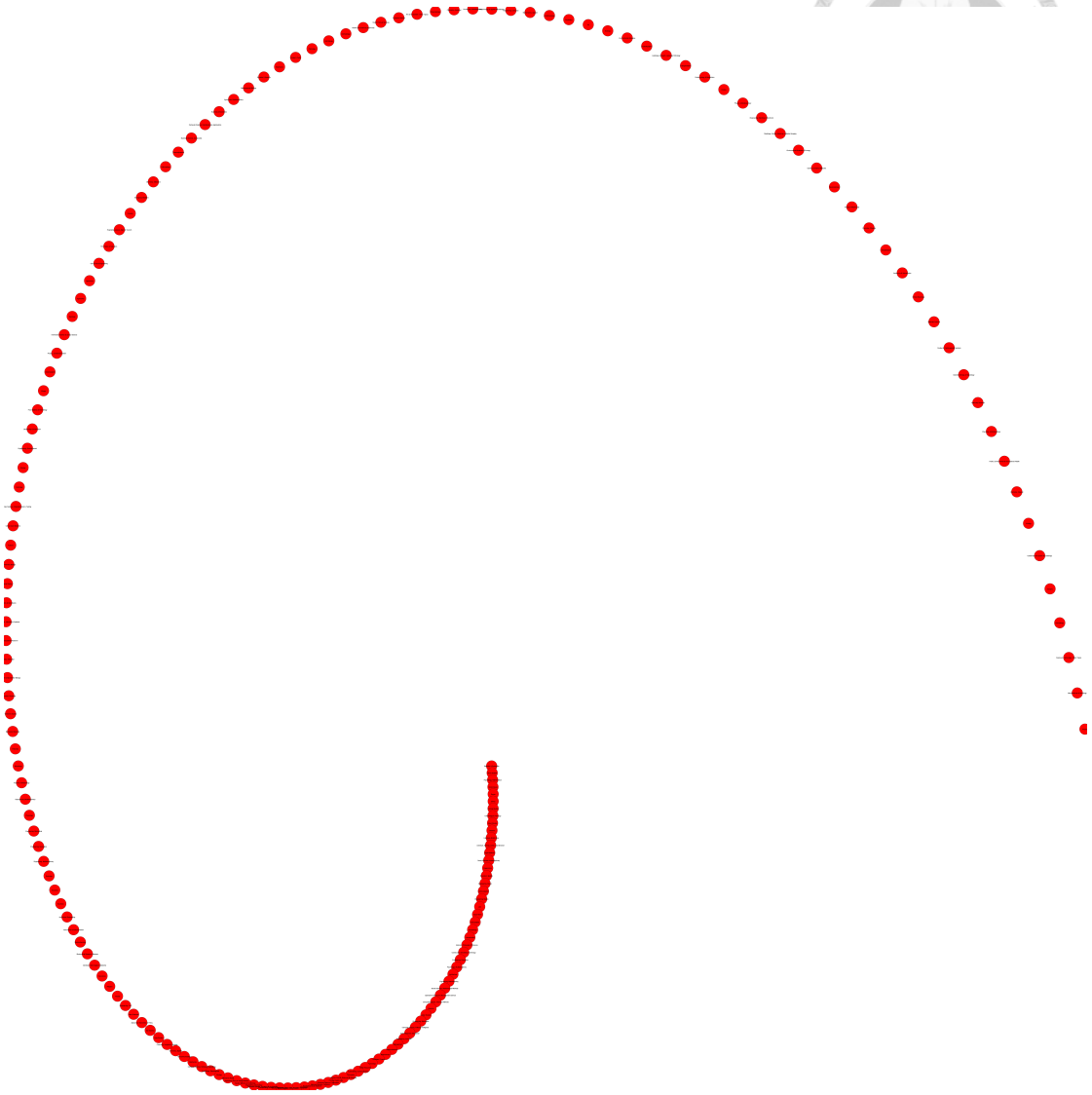


Figure 12: Similar subjects graph for Literature, Statistics Probability

doi:10.6342/NTU202501640

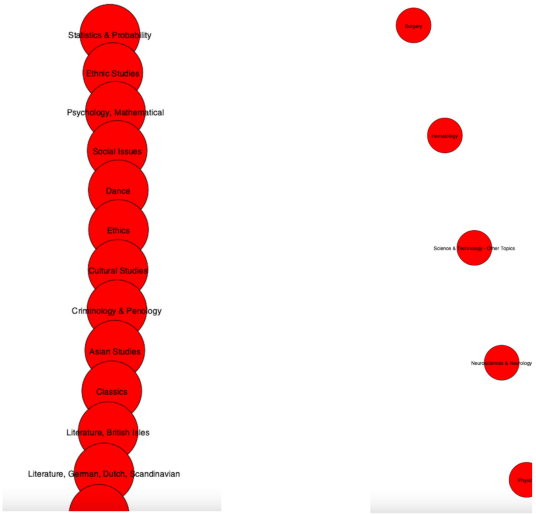


Figure 13: the most similar subjects (left) and the least similar subjects(right) with Statistics Probability

6 References

1. Abbasi, A., Chung, K.S.K., Hossain, L.: Egocentric analysis of co-authorship network structure, position and performance. Information Processing & Management, 48(4), 671-679 (2012).
2. A. Velez-Estevez, P. García-Sánchez, J. A. Moral-Munoz, M. J. Cobo: Why do papers from international collaborations get more citations? A bibliometric analysis of Library and Information Science papers. Scientometrics 127:7517–7555 (2022).
3. Matteo Cinelli, Giovanna Ferraro, Antonio Iovanella: Connections matter: a proxy measure for evaluating network membership with an application to the Seventh Research Framework Programme. Scientometrics 127:3959–3976 (2022)
4. Bos, J.: Numerical optimization of the thickness distribution of three-dimensional structures with respect to their structural acoustic properties. Structural and Multidisciplinary Optimization, 32(1), 12-30 (2006).
5. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of networks with aging of sites. Physical Review E, 62(2), 1842 (2012).
6. Jung, H., Phoa, F.K.H., Ashouri, M. A Leading Author Model for the Popularity Effect on Scientific Collaboration. In: Benito, R.M., Cherifi, C., Cherifi, H., Moro, E., Rocha, L.M., Sales-Pardo, M. (eds) Complex Networks Their Applications X. COMPLEX NETWORKS 2021. Studies in Computational Intelligence, vol 1015. Springer, Cham.(2022)
7. Ghiasi, G., Harsh, M., Schiffauerova, A.: Inequality and collaboration patterns in Canadian nanotechnology: implications for pro-poor and gender-inclusive policy. Scientometrics, $115(2), 785–815$ (2018).

doi:10.6342/NTU202501640

8. Kuppler, M. Predicting the future impact of Computer Science researchers: Is there a gender bias?. Scientometrics 127, 6695–6732 (2022)

9. Jeong, H., Neda, Z., Barabasi, A-L.: Measuring preferential attachment in evolving networks. Europhysics Letters, 61(4), 567 (2003).

10. Jung, H., Lee, J-G., Kim, S-H.: On the analysis of fitness change: fitness-popularity dynamic network model with varying fitness. Journal of Statistical Mechanics: Theory and Experiment, 2020(4), 043407 (2020).

11. Jung, H., Lee, J-G., Lee, N., Kim, S-H.: PTEM: A popularity-based topical expertise model for community question answering. Annals of Applied Statistics, 14(3), 1304-1325 (2020).

12. Louis, T.A.: Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 44(2), 226-233 (1982).

13. Lu, H., Feng, Y.: A measure of authors' centrality in co-authorship networks based on the distribution of collaborative relationships. Scientometrics, 81(2), 499-511 (2009).

14. Jung, H., Phoa, F.K.H., Ashouri, M.: A Leading Author Model for the Popularity Effect on Scientific Collaboration. Complex Networks Their Applications X. COMPLEX NETWORKS 2021. Studies in Computational Intelligence, vol 1072. Springer(2022)

15. Merton, R.K.: The Matthew effect in science: The reward and communication systems of science are considered. Science, 159(3810), 56-63 (1968).

16. Metz, T., Jackle, S.: Patterns of publishing in political science journals: An overview of our profession using bibliographic data and a co-authorship network. PS, Political Science & Politics, 50(1), 157-165 (2017).

17. Perc, M.: The Matthew effect in empirical data. Journal of The Royal Society Interface, 11(98), 20140178 (2014).

18. Rode, S.M., Pennisi, P.R.C., Beaini, T.L.,Curi, J.P., Cardoso, S.V., Paranhos, L.R.: Authorship, plagiarism, and copyright transfer in the scientific universe. Clinics, 74 , 1312 (2019).

19. Roy, S., Ravindran, B.: Measuring network centrality using hypergraphs. Proceedings of the Second ACM IKDD Conference on Data Sciences, (pp. 59-68) (2015).

doi:10.6342/NTU202501640