Journal of Informetrics 18 (2024) 101486

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Dependency, reciprocity, and informal mentorship in predicting long-term research collaboration: A co-authorship matrix-based multivariate time series analysis

Yongjun Zhu [a], Donghun Kim [a], Ting Jiang [a], Yi Zhao [b], Jiangen He [c], Xinyi Chen [d], Wen Lou [e,*]

[a] Department of Library and Information Science, Yonsei University, Seoul, South Korea
[b] School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China
[c] School of Information Sciences, The University of Tennessee, Knoxville, United States
[d] Department of Cultural Media, Yonsei University, Seoul, South Korea
[e] School of Economics and Management, East China Normal University, Shanghai, China

ARTICLE INFO

Keywords:
Long-term research collaboration
Co-authorship prediction
dependency
Reciprocity
Informal mentorship
Interpretable machine learning

ABSTRACT

In this study, we examine the roles of dependency, reciprocity, and informal mentorship in the prediction of long-term research collaboration in five disciplines. We use co-authorship matrix-based multivariate time series features and interpretable machine learning to train long-term collaboration prediction models and interpret the feature importance of trained models. Overall, long-term research collaboration that is defined using various standards was rare across the examined disciplines, and the prediction results were moderate to good. We found dependency, reciprocity, and informal mentorship to have different roles in different disciplines. Among the three, informal mentorship was important in predicting long-term research collaboration in Agriculture, Geology, and Library and Information Science. Reciprocity, which measures the interdependence between two researchers was important to prediction in the fields of Agriculture and Geology. Finally, dependency was important in all the disciplines with varying degrees of importance.

1. Introduction

Collaboration among individual researchers has traditionally been encouraged in the scientific community (Bozeman et al., 2013). The benefits of collaboration are diverse, including opportunities to use external resources, divide labor, and alleviate academic isolation, among many others (Fox & Faver, 1984). Although collaboration comes with a cost such as time for exchange and possible delay of project, collaboration is widespread among researchers (Traoré & Landry, 1997). For these reasons, governments and institutions have launched numerous initiatives to develop and support collaboration to improve research quality (Melin, 2000), efficiency (Fox & Faver, 1984), and productivity (Lee & Bozeman, 2005a). Specifically, long-term research collaboration is associated with innovation (Creamer, 2004).

During collaboration, researchers depend on each other to a certain degree. For example, one may depend on another to pool her

* Corresponding author.
E-mail address: wlou@infor.ecnu.edu.cn (W. Lou).

https://doi.org/10.1016/j.joi.2023.101486
Received 11 January 2023; Received in revised form 14 December 2023; Accepted 17 December 2023
Available online 22 December 2023

knowledge with others, overcome a lack of resources (e.g., instrumentation), or increase her scientific popularity/visibility (Katz & Martin, 1997). Dependency does not necessarily mean that someone does not have the ability to work alone, and in many cases, researchers decide to work together and depend on others for efficiency. Everyone in a team can take charge of the part that matches her own specialty and thus the labor can be divided. We believe dependency encourages collaboration, and collaboration creates dependency, hence dependency plays a positive role in long-term research collaboration. In the same vein, dependency does not form in casual collaboration, and casual collaboration does not often develop into a long-term collaborative relationship.

Before collaboration, one who initiates or leads a work invites another to work together. An invitation can be reciprocal, meaning that two researchers invite each other to their respective works, or an invitation can be one-sided, in which one never or rarely invites the other to her work but frequently joins the other's work. The latter may happen more frequently in junior–senior collaboration than in collaboration between peers because junior researchers may need advice and resources from senior researchers. It is also possible that junior researchers who are experts in emerging fields receive many invitations from both junior and senior researchers to work together. While one-sided collaboration can be long-lasting, we believe this type of collaborative relationship is more difficult to maintain than the reciprocal relationship. One-sided collaboration can be interpreted as "I can help you when you need me, but I don't need your help"; a relationship is easier to maintain when both people need each other. Accordingly, we believe reciprocity is another key factor that affects long-term research collaboration.

Mentorship exists in academia. A representative example is that of advisers and doctoral students. Doctoral students obtain multiple years of training and supervision from their advisers and they naturally become each other's collaborators. In this adviser–advisee collaborative relationship, students may work on their own or on adviser-assigned research work, and advisers offer mentoring. It is also possible that advisors pursue their own research work and students help advisers with tasks such as data collection, during which no mentoring occurs. This collaborative relationship may be maintained over later years and can be developed into a new collaboration type—junior–senior collaboration—after students successfully grow as independent researchers. However, mentorship is not solely a property of adviser–advisee collaboration; it also exists in collaboration between peers and between researchers of unequal ranks. AlShebli et al., (2020)) used the term "informal mentorship" to describe the mentoring relationship between junior and senior researchers without formal supervisory roles using coauthorship data. Mentoring motivates and fosters collaboration, and this is important in long-term research collaboration.

Although co-authorship is a partial indicator of collaboration, it has widely been used as an effective tool for understanding research collaboration (Bozeman et al., 2013). Co-authorship is by no means a perfect measure of collaboration, yet it is invariant, verifiable, inexpensive, and practical (Katz & Martin, 1997). In this study, we use co-authorship as a proxy measure to investigate how dependency, reciprocity, and informal mentorship affect long-term research collaboration. To achieve this goal, we aim to address the following research questions.

- RQ1: How do we incorporate co-authorship types and the temporal nature of collaboration into a co-authorship-based framework that can be used to predict long-term research collaboration?
- RQ2: How important are dependency, reciprocity, and informal mentorship in predicting long-term research collaboration?

The rest of this paper is organized as follows. In related work, we discuss studies on research collaboration and co-authorship followed by the methods and data section, where we introduce the datasets, the concept of the co-authorship matrix, and the time series analysis methods for the prediction of long-term research collaboration. Following that, we present the experimental results and discuss findings and limitations. Finally, we conclude the paper.

1.2. Related work

Research collaboration has been explored in a number of studies along with the growing importance of collaborative research (Wray, 2006). Co-authorship is the most common measure of research collaboration (Savanur & Srikanth, 2010) and sub-authorship, such as mentions in the acknowledgements, has been also suggested as a measure (Cronin et al., 2003, 2004). Although co-authorship is a partial indicator of collaboration (Katz & Martin, 1997) and not every research collaboration will lead to a publication (Bukvova, 2010), it has widely been used as an effective tool for understanding research collaboration. Co-authorship has been used to investigate spatial patterns of research collaboration (Hoekman et al., 2010; Yarime et al., 2009), interinstitutional and international research collaboration (Cheng et al., 2014), influential authors and groups in research collaboration, university–industry research collaboration (Abramo et al., 2011), the relationship between research collaboration and research productivity (Abramo et al., 2017; Lee & Bozeman, 2005b; Lee et al., 2012; Levitt & Thelwall, 2016; Ye et al., 2011), and citation patterns (Gazni & Thelwall, 2014).

The factors that may have impact on research collaboration have been studied. Transdisciplinarity between authors (Bu et al., 2018), authors' scientific impact (Amjad et al., 2017), and research team size (Larivière et al., 2015) have been studied. Studies have shown that high-impact researchers tend to collaborate with researchers who have expertise in diverse research topics (Bu et al., 2018). Junior researchers who collaborate with well-known senior researchers in their early career stage have a higher chance of success (Amjad et al., 2017). Top researchers tend to have a higher propensity to collaborate internationally than their lower-performing colleagues, and their productivity is closely related to international research collaboration (Abramo et al., 2019). It has been reported that an increase in team size is positively related to an increase in impact (Larivière et al., 2015). Few studies have specifically explored long-term research collaboration. Bu et al. (Bu et al., 2018) investigated the relationship between persistent research collaboration and publication quality using co-authorship data. They found that transdisciplinarity and diversity in coauthor's scientific age and impact had a positive impact on publication quality, whereas researchers who persistently worked in large groups

tended to publish lower-impact papers.

Studies on the predication of long-term research collaboration are relatively few. A study aiming to predict future co-authorship for junior researchers found that affiliation overlap, geographic distance, and research topic similarity had been important features for prediction (Tsai & Lin, 2016). Wang et al. (2019) posited that researchers who attend the same conference may well collaborate in the future and this factor turns out to be effective in predicting the duration of collaboration. Drawing on a global dataset of more than three million papers, Shen et al. (2022) adopted regression model to quantify the relationship between gender composition and collaboration continuity. The results showed that intra-gender collaboration could persist longer than inter-gender collaboration. Based on text and network structure information, Wang et al. (2021) learned vector representation of each scholar and then incorporated additional author attributes such as demographics, research, influence, and sociability to predict life-time collaborators for early-stage researchers. Their findings suggested that author attributes had been useful to predict life-time collaborators.

Different from the reviewed studies, we aim to incorporate co-authorship types and the temporal nature of collaboration into a co-authorship-based framework and predict long-term research collaboration. In addition, we further explore the role of dependency, reciprocity, and informal mentorship in predicting long-term research collaboration in representative scientific disciplines.

2. Methods

2.1. Coauthorship matrix

As the base framework for understanding dependency, reciprocity, and informal mentorship in long-term research collaboration, we propose the co-authorship matrix (see Fig. 1). From two researchers' coauthored papers, we compute the number of each collaboration type. For example, A_fB_{co} denotes the number of coauthored papers where researcher A is the first author and researcher B is one of the coauthors. By excluding the cases of two or more first authors in a paper, which is rare (Lapidow & Scudder, 2019) and sometimes considered as unethical (Agoramoorthy, 2017), we have eight co-authorship types.

Dependency is a temporal concept. In a long-term research collaboration, we estimate the degree of dependency by considering how many times and how often a researcher invites another to coauthor a paper where the inviting author is the lead author (i.e., the first or corresponding author). For example, a set of yearly A_fB_{co} and $A_{cor}B_{co}$ can be used to estimate how much researcher A depends on researcher B. It should be noted that if researcher A frequently invites researcher B as a coauthor to conduct research, although what kind of resources or knowledge that researcher B can provide to researcher A is unclear, it implies a significant value of researcher B to researcher A. Given that the first and corresponding authors lead a research project and are generally responsible for constructing teams by inviting others, A_fB_{co} and $A_{cor}B_{co}$ are two important co-authorship types to estimate dependency.

Reciprocity can be interpreted as "give and take". Invitation to co-authorship can be one-sided, meaning that one never or rarely invites the other to coauthor but frequently joins the other's work. We compare A_fB_{co} and $A_{cor}B_{co}$ with $A_{co}B_f$ and $A_{co}B_{cor}$ to estimate the degree of reciprocity. A good balance of "give and take" means two authors both need each other, and this may have positive impact on their long-term collaboration.

If there exists a strong informal mentorship between two researchers, regardless of the degree of reciprocity, the collaborative relationship may be long-lasting. In this study, we aimed to explore informal mentorship by only taking into account the relationship between the first and corresponding authors while AlShebli et al. (2020) also considered coauthors in their study. Because we cannot tell, from the coauthorship data, whether there is a formal supervising role between the first and corresponding authors or not, we use the term "informal mentorship" in the study. A_fB_{cor} is used to estimate the degree of informal mentorship of B on A.

Finally, two researchers may also collaborate with each other in another researcher's project, which is captured by $A_{co}B_{co}$ and $A_{cor}B_{cor}$. Two corresponding authors in a paper may be a rare case but it does exist. For example, a researcher may have two advisers who supervise the work together. A high $A_{co}B_{co}$ means the two researchers have many common collaborators, which may facilitate their direct collaboration.

		Researcher B		
		first author	coauthor	corresponding author
Researcher A	first author	NA	A_fB_{co}	A_fB_{cor}
	coauthor	$A_{co}B_f$	$A_{co}B_{co}$	$A_{co}B_{cor}$
	corresponding author	$A_{cor}B_f$	$A_{cor}B_{co}$	$A_{cor}B_{cor}$

Fig. 1. Co-authorship matrix.

2.2. Multivariate time series analysis

The co-authorship matrix shown in Fig. 1 is a static capture of collaboration at a certain time point (e.g., a certain year). From the perspective of long-term collaboration, dependency, reciprocity, and informal mentorship are interrelated and may affect each other. Between two researchers, all the three types of relationship may co-exist and one type of relationship may completely or partially transition to another type. This nature motivated us to consider the time dimension to investigate the dynamic interplay among the three in modeling long-term research collaboration. In investigating two researchers' long-term collaboration, we consider an ordered set of these matrices to capture the temporal aspect. The co-authorship tensor shown in Fig. 2 is one such data structure that combines n-year co-authorship matrices. For an in-depth investigation of the temporal evolution of two researchers' collaboration, a co-authorship tensor can be transformed into the time series co-authorship matrix shown in Fig. 2, in which a row represents collaboration between two researchers within a year. We generate features from a raw time series co-authorship matrix by extracting new features that represent important traits, trends, and patterns in the original time series data. Given a line that represents the temporal change of each co-authorship type, we extracted the maximum, minimum, median, mean, and number of peaks of the line. After the features are generated, time series prediction and interpretable machine learning methods can be applied.

2.3. Machine learning for time series

To predict long-term research collaboration, we use data from previous years and predict future collaboration by investigating trends and patterns during the collaboration in previous years. Specifically, given two time periods, we train a prediction model utilizing data of the first period and predict the occurrence of continuing collaboration in the second period. This task can be reframed as a binary classification, as given a pair of researchers, a trained model gets collaboration data between the two in the first time period as input and produce the probability that the two continue to collaborate in the second period.

2.4. Data

Philosophy, Agriculture, Geology, Sociology, and Information Science & Library Science (LIS) were chosen as examples of the five broad categories (i.e., Arts & Humanities, Life Sciences & Biomedicine, Physical Sciences, Social Sciences, and Technology) used by Web of Science to classify research areas. We collected 1,035,024 bibliographic records of all the document types having address information that were published between 2008 and 2020 in the Web of Science Core Collection. 2008 was chosen because Web of Science began to link authors and addresses in 2008. Collected metadata included Web of Science UID, authorship information (including author names and addresses), reprint information (including the name and address of corresponding author) and publication year. For a single bibliographic record, we regarded the first person in the authorship information as the first author, and the person (might be multiple) who is listed in reprint information as the corresponding author. For each discipline, Fig. 3 shows the distribution of papers having different numbers of authors. Overall, more than 80 % of papers in Philosophy were single-authored, whereas more than 90 % of papers in Agriculture and Geology were written by multiple authors. Sociology and LIS had a similar proportion of single- and multiple-authored papers.

During author name disambiguation, we used approximate string matching to allow for various writing conventions and typos to

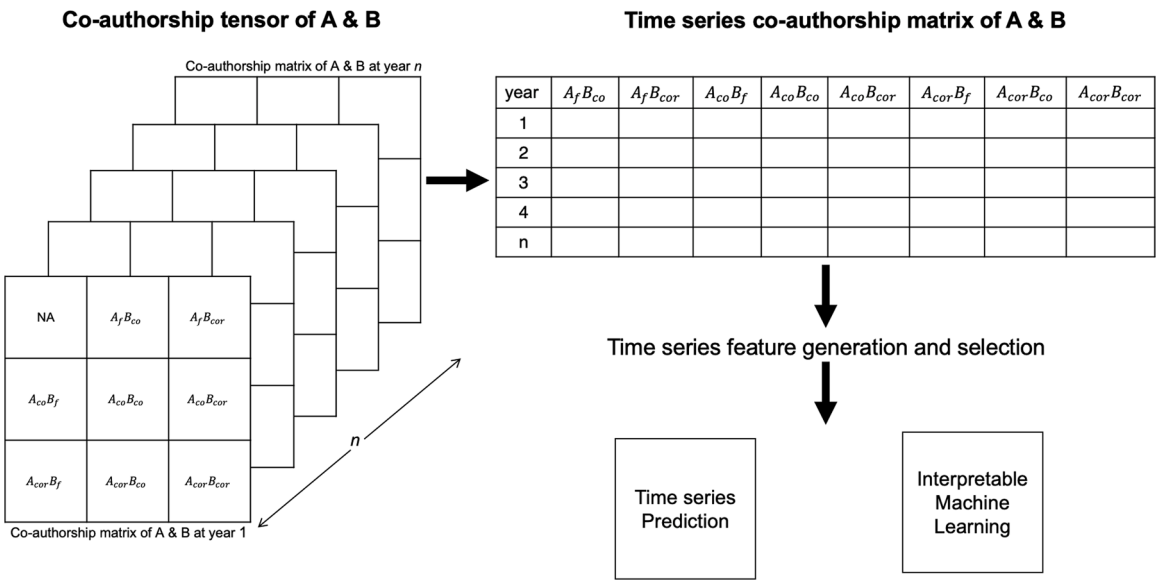

Fig. 2. Time series co-authorship matrix.

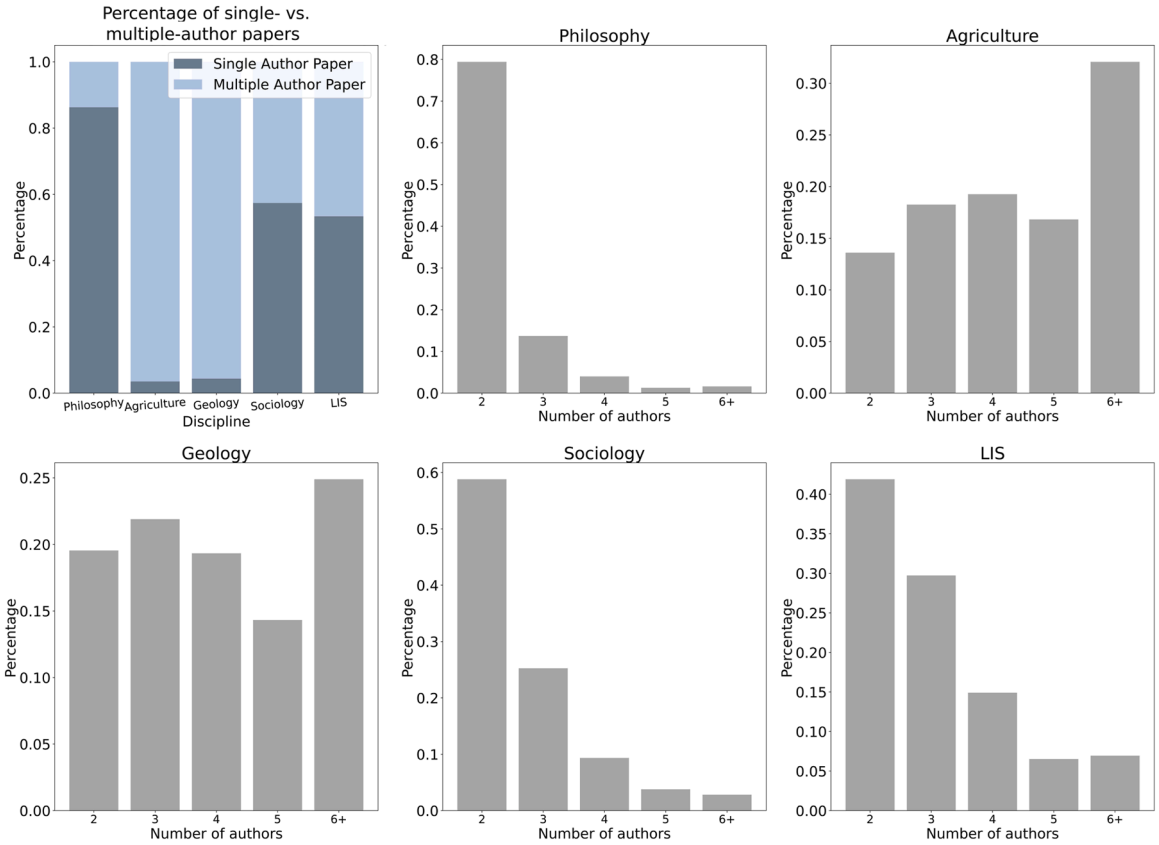

Fig. 3. Distribution of single- and multiple-authored papers in five disciplines.

minimize false negatives. The string similarity score of the affiliations of any two authors with the same name were compared to determine whether the two authors were the same person. To select a proper threshold of similarity score above which we can decide with a certain confidence that the two authors are the same person, we random sampled ten pairs for every 0.1 similarity score interval from 0 to 1, and we manually checked the percentage of true cases. For pairs in which authors had multiple addresses, we computed the similarity score for each address pair and used the maximum value as the similarity score of the two authors. We found that when similarity scores were higher than 0.8, the percentage of true cases was 100 %. Therefore, we randomly sampled another 100 pairs having similarity scores in the range between 0.7 and 0.8, and we found that nine out of ten pairs were true cases when the score was higher than 0.7. To minimize deletion rate and false negatives, we adopted 0.7 instead of 0.8 as the similarity score threshold. Table 1 shows the final data after author name disambiguation.

3. Results

A long-term research collaboration is a collaboration that continues multiple years after the first collaboration between two researchers. The prediction task can be formulated as predicting collaboration in the second period given the coauthorship history of two researchers in the first period. Given that the collected data were dated between 2008 and 2020, we tested different combinations of periods, where the first period was either five or seven years, and the second period was three, five, or seven years. The three-year period was not considered as an option for the first period because we thought three years was not a long enough time to observe patterns in two researchers' long-term collaboration. If two researchers who collaborated in the first period also collaborated in the

Table 1
Data description after author name disambiguation.

Category	No. of papers	No. of unique authors	No. of author pairs
Philosophy	75,331	50,058	20,552
Agriculture	439,643	1,109,220	4,288,345
Geology	318,025	704,540	3,201,435
Sociology	97,914	108,752	110,661
LIS	104,111	119,733	220,708

second period, we consider that the two have a long-term collaboration relationship. Table 2 shows the number of total author pairs and long-term collaboration pairs in different settings. We can see that long-term research collaboration is extremely rare in all five disciplines.

Given the sequences of collaboration history in the training data (i.e., data for the first period), we extracted time series features using Python library's *tsfresh* package (Christ, et al., 2018). Specifically, given a time series co-authorship matrix (Fig. 2), *maximum, minimum, median, mean*, and *number of peaks* were extracted for each of the eight co-authorship types, which resulted a total of 40 features for each collaboration pair.

Our data were unbalanced and had a small proportion of long-term collaboration pairs that we intended to predict. To solve the problem, we undersampled negative cases (i.e., the author pairs that were not long-term research collaborations) in each discipline to match the number of positive and negative cases. After preparing the training and test datasets, we engineered features appropriately by (1) dropping features that were either missing in more than 70 % cases or had zero values in more than 85 % cases; (2) min-max normalizing the features; and (3) selecting the best-performing feature sets by sequentially removing redundant features using cross-validation and the F1 score as the scoring parameter. Tree-based algorithms, including scikit-learn *Random Forest, Gradient Boosting Classifier* (Pedregosa et al., 2011), and *XGBoost* (Chen & Guestrin, 2016) were used to train models using five-fold cross validation. We report the prediction results for *XGBoost*, which showed better performance than *Random Forest* and *Gradient Boosting Classifier* with mean F1 scores of 0.74, 0.73, and 0.72 respectively. After testing all the combinations of collaboration periods and sequential feature selection, we report the best models in Table 3.

As shown in Table 3, all the best models were trained on a collaboration history of seven years. Because our data were dated between 2008 and 2020 and not every collaboration began in 2008, given the seven years in the first period, we tested five or three years within the second period. Overall, we obtained F1 scores between 0.6 and 0.8, and superior performance was observed for Agriculture, Geology, and LIS.

While the features of the models in Table 3 were all derived from the relationship between pairs of researchers, we additionally considered features of individual researchers to understand their roles in prediction. Specifically, we added two features about individual researchers including the number of papers and the number of coauthors of each researcher, computed from the training dataset. Table 4 shows the results.

Considering productivity (i.e., the number of papers published by each of the collaborating researcher in the period of training data) improved model performance slightly in four disciplines with F1 scores increased ranging from 0.0093 in Agriculture to 0.0516 in Philosophy. We saw worsening performance in LIS. Similarly, adding the number of coauthors improved F1 in three disciplines and worsened F1 in two disciplines. Geology had the biggest performance increase of 0.0176 while Philosophy had the biggest performance decrease of 0.0748. Adding both features increased F1 score in Sociology by 0.1. In summary, we found that adding features of individual researchers did not necessarily improve model performance and their roles varied across disciplines. One of many possible reasons is that features derived from the relationship between two researchers and individual researchers are correlated and may overlap. This may increase or decrease model performance.

Based on the results of Table 3, the relative contributions of features to the prediction task were obtained from feature importance scores, which were computed by the built-in *XGBoost* algorithm. Fig. 4 lists the selected features in Table 3 in descending order.

Table 2
Number of total and long-term collaboration author pairs in each period.

	first period	second period	No. of total author pairs	No. of long-term author pairs
Philosophy	5	3	5,379	78
	5	5	3,138	62
	5	7	1,488	32
	7	3	3,138	36
	7	5	1,488	19
Agriculture	5	3	16,9162	2,815
	5	5	95,717	1,881
	5	7	37,295	952
	7	3	95,717	917
	7	5	37,295	555
Geology	5	3	128,552	2,808
	5	5	79,780	2,471
	5	7	37,463	1,523
	7	3	79,780	1,311
	7	5	37,463	955
Sociology	5	3	36,623	475
	5	5	21,379	417
	5	7	9,963	229
	7	3	21,379	217
	7	5	9,963	135
LIS	5	3	822,44	1,003
	5	5	50,019	814
	5	7	27,562	456
	7	3	50,019	336
	7	5	27,562	229

Table 3
Results of best models for long-term research collaboration prediction.

Discipline	First period	Second period	No. of features	accuracy	precision	recall	F1	ROC AUC
Philosophy	7	5	7	0.7393	0.9000	0.5667	0.6762	0.7333
Agriculture	7	5	18	0.7755	0.8290	0.6943	0.7556	0.7755
Geology	7	5	24	0.7549	0.8287	0.6428	0.7239	0.7549
Sociology	7	3	14	0.6681	0.7452	0.5205	0.6076	0.6681
LIS	7	3	16	0.7926	0.8528	0.7118	0.7746	0.7926

Table 4
Results of models after adding features of individual researchers.

Discipline	Original		Original + No. of papers		Original + No. of coauthors		Original + No. of papers + No. of coauthors	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Philosophy	0.6762	0.7333	0.7278	0.7500	0.6014	0.7000	0.6719	0.7000
Agriculture	0.7556	0.7755	0.7649	0.7792	0.7591	0.7741	0.7657	0.7785
Geology	0.7239	0.7549	0.7539	0.7686	0.7415	0.7601	0.7591	0.7698
Sociology	0.6076	0.6681	0.6357	0.6635	0.6139	0.6377	0.7076	0.7250
LIS	0.7746	0.7926	0.7670	0.7770	0.7603	0.7619	0.7603	0.7643

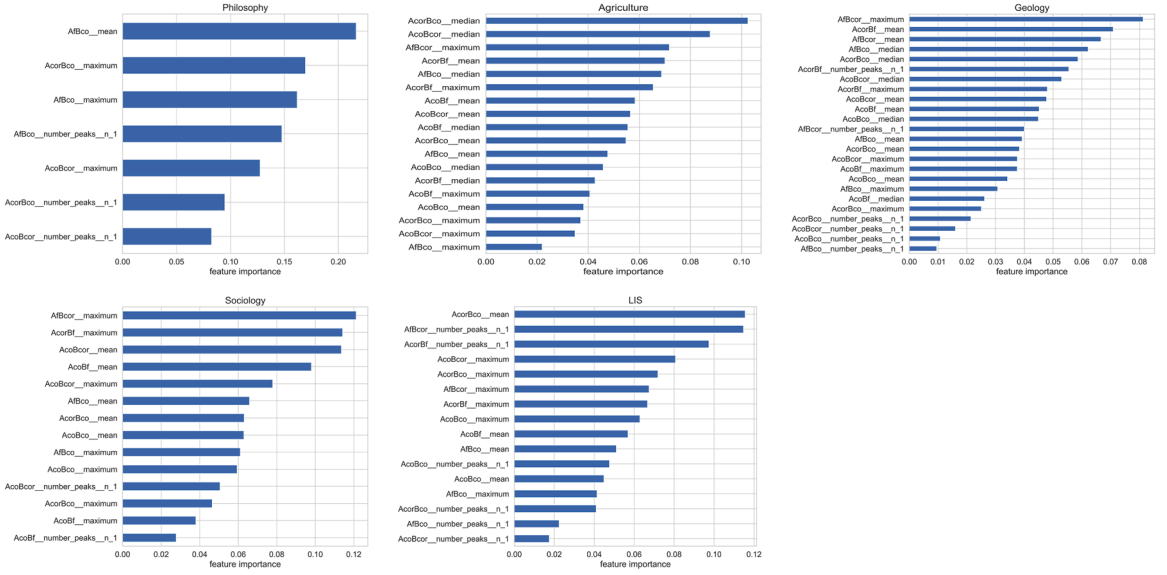

Fig. 4. Important features in the prediction of long-term collaboration in five disciplines.

Overall, the maximum, mean, and median of various co-authorship types were in the list of top five important features, and this denotes that intensity of a single layer (i.e., maximum) and the average intensity over years (i.e., mean and median) are important characteristics of long-term research collaboration. To better understand the importance of each co-authorship type, we merged the features by their co-authorship types. For example, we calculated importance of A_fB_{co} by adding the importance of all features that started with A_fB_{co}, such as A_fB_{co}_max, A_fB_{co}_min, A_fB_{co}_median, etc. By doing so, we obtained the temporal feature importance of each co-authorship type in each discipline, as shown in Fig. 5.

As shown in Fig. 5, $A_{cor}B_{cor}$ was not important to the prediction. There may be multiple reasons behind this result. One of the possible reasons may be its relatively underrepresented nature in the publication data. A research publication with two or more corresponding authors are not common compared with other types of co-authorship types. $A_{co}B_{co}$ was the most important in LIS, followed by Sociology, Geology, and Agriculture. Thus, we can see that in LIS, jointly participating in a third party's research projects is an important signal of long-term research collaboration between two researchers. Overall, except for Philosophy, which showed only three co-authorship types (i.e., A_fB_{co}, $A_{co}B_{cor}$, and $A_{cor}B_{co}$), the other disciplines showed various levels of feature importance in seven coauthorship types. To understand the roles of dependency, reciprocity, and informal mentorship, we merged the seven co-authorship types into three, and we examined the characteristics of each discipline, as shown in Table 5. When we created co-authorship pairs, we set A as the researcher with more papers and B as the one with fewer papers between the two. Therefore, in Table 5, we show two versions, one for each perspective: the more experienced researcher and the less experienced one.

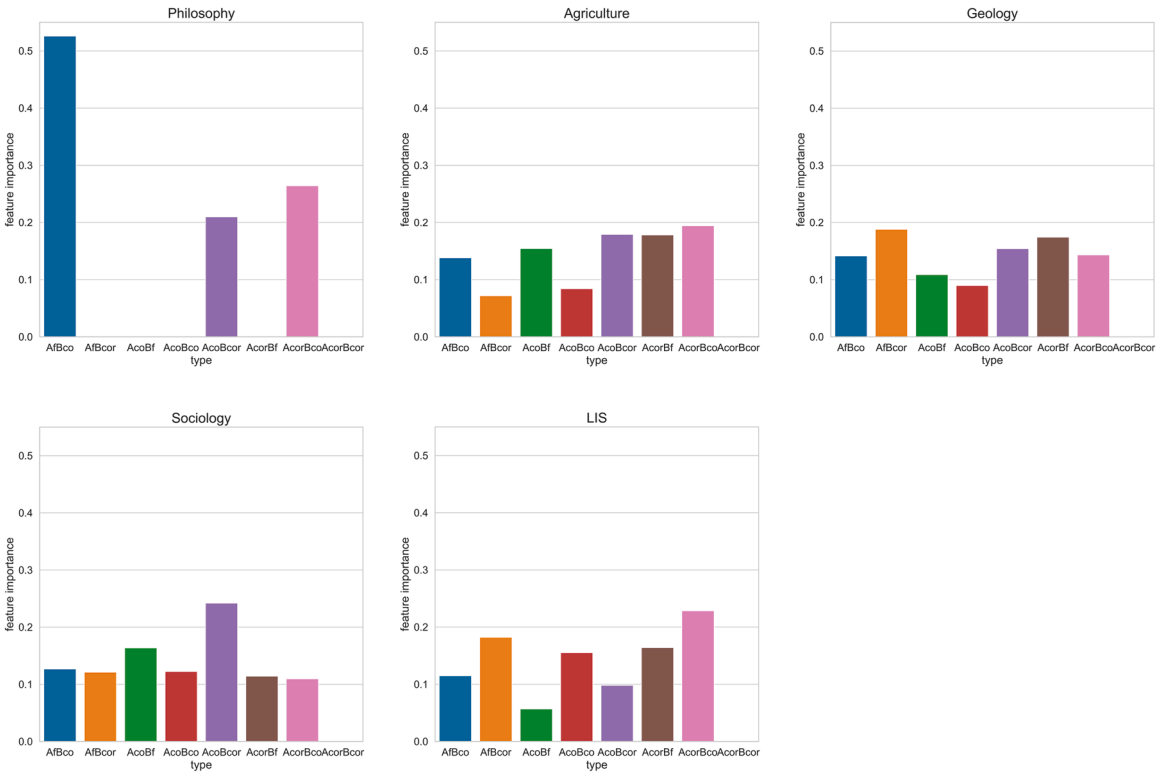

Fig. 5. Temporal feature importance of each co-authorship type.

Table 5
Importance of dependency, reciprocity, and informal mentorship in five disciplines.

Discipline	Dependency				Reciprocity		Informal mentorship	
	1 A_fB_{co}	2 $A_{co}B_f$	3 $A_{cor}B_{co}$	4 $A_{co}B_{cor}$	1 $\frac{dependency\ 1+3}{dependency\ 2+4}$	2 $\frac{dependency\ 2+4}{dependency\ 1+3}$	1 A_fB_{cor}	2 $A_{cor}B_f$
Philosophy	**high (0.526)**	low (0.000)	**High (0.264)**	**High (0.210)**	Low (3.765)	Low (0.266)	low (0.000)	low (0.000)
Agriculture	medium (0.138)	**high (0.154)**	**high (0.194)**	**high (0.179)**	**high (0.997)**	**high (1.002)**	medium (0.072)	**high (0.178)**
Geology	medium (0.142)	medium (0.109)	medium (0.143)	**high (0.154)**	**high (1.083)**	**high (0.923)**	**high (0.188)**	**high (0.174)**
Sociology	medium (0.127)	**high (0.164)**	medium (0.110)	**high (0.242)**	medium (0.583)	medium (1.716)	medium (0.121)	medium (0.114)
LIS	medium (0.115)	medium (0.057)	**high (0.228)**	medium (0.098)	medium (2.214)	medium (0.452)	**high (0.182)**	**high (0.164)**

For dependency and informal mentorship: low: value <= 0.05, medium: 0.05 < value <= 0.15, high: value >0.15; For reciprocity: low: value <= 0.3 or value >= 3.33, medium: 0.3 < value <= 0.7 or 1.43 <= value < 3.33, high: value > 0.7 or value < 1.43. High values are highlighted.

In Table 5, high dependency (i.e., high importance scores for dependency 1, 2, 3 and 4) denotes that the degree to which one researcher depends on the other, manifested by inviting her to be a coauthor when the inviting researcher is the first or corresponding author of the co-authoring paper, is important in predicting long-term research collaboration. Dependency 1 and 3 measure how dependent A, the more experienced researcher, is on B, and dependency 2 and 4 measures how dependent B, the less experienced researcher, is on A. Reciprocity measures whether the two types of dependency are balanced or not, where values close to 1 denote a perfect balance and values either too small or too large denote an extreme unbalance. Two versions of reciprocity are possible from the perspectives of experienced or less experienced researchers. High reciprocity (i.e., importance scores close to 1 in reciprocity 1 and 2) denotes a two-way dependency between two researchers is equally important in prediction. In other words, "give and take" is important in predicting long-term research collaboration. Finally, high informal mentorship (i.e., high importance scores in informal mentorship 1 and 2) denotes that the collaborative relationship between advisee and adviser, manifested by the relationship between the first and corresponding authors, is important in predicting long-term research collaboration.

Based on the above explanation, we discuss the characteristics of each discipline as follows. Overall, Philosophy showed high dependency (i.e., for dependency 1, 3, and 4), low reciprocity, and low informal mentorship, denoting that dependency had an important role in predicting long-term research collaboration in the discipline, whereas reciprocity and informal mentorship were not that important. Given that researchers in the field rarely coauthor with others, as shown in Fig. 3, they may invite others to coauthor a paper only when they have a strong need for the invited researchers' expertise, and thus, dependency is a strong signal of long-term research collaboration in the discipline. Specifically, the importance score for dependency 1 is extremely high, which can be interpreted as meaning that the dependency of more experienced researchers on less experienced ones is more important in prediction. Agriculture showed high dependency (i.e., for dependency 2, 3, and 4), high reciprocity (i.e., for reciprocity 1 and 2), and a high score for informal mentorship 2, denoting that given a high collaboration rate in the discipline, all three concepts are important to predicting long-term research collaboration. Specifically, the high score in informal mentorship 2, but not in informal mentorship 1, denotes that the advisee-adviser relationship is a strong signal of long-term research collaboration in the discipline. Geology showed a high dependency for only one co-authorship type (i.e., dependency 4), high reciprocity (i.e., for reciprocity 1 and 2), and high informal mentorship (i.e., for informal mentorship 1 and 2). Compared with Philosophy and Agriculture, Geology had a lower level of dependency, denoting that dependency is less important in predicting long-term research collaboration in the discipline. Specifically, the degree to which less experienced researchers invite more experienced researchers to papers they supervise is an important signal in prediction. In addition, two types of informal mentorships are equally important in Table 5, and this denotes that the traditional informal mentorship relationship between junior and senior researchers is not the only type of informal mentorship in the discipline, and in turn, various types of informal mentorship are important to prediction. Sociology showed high scores for dependency 2 and 4, medium reciprocity, and informal mentorship. From these results, it seems that the dependency of less experienced researchers on more experienced ones when the inviting authors are either the first or the corresponding authors is an important signal for prediction. Finally, LIS showed high dependency for only one co-authorship type (i.e., dependency 3), medium reciprocity, and high informal mentorship. Different from Geology, the degree to which more experienced researchers invite less experienced researchers to papers they supervise is an important signal in prediction. The two disciplines showed the same pattern in terms of informal mentorship where both directions are important.

4. Discussion

In this study, we aimed to understand the roles of dependency, reciprocity, and informal mentorship in long-term research collaboration through prediction models. Specifically, we (1) proposed a co-authorship matrix that represents multiple co-authorship types; (2) introduced a temporal co-authorship matrix that incorporates the time dimension of collaboration; (3) extracted time series features from the temporal co-authorship matrix; (4) trained collaboration prediction models; and (5) interpreted the models by discussing the feature importance of each co-authorship type. Overall, long-term research collaboration as defined using various standards is rare across all the examined disciplines, and the prediction results were moderate to good. Despite its various contribution to the understanding of long-term research collaboration, this study has a few limitations. First and foremost, co-authorship data is a partial representation of research collaboration, and lasting collaboration that failed to produce research publications were not examined. Second, dependency, reciprocity, and informal mentorship were operationalized using co-authorship data although they have much broader impact that cannot be fully captured by publication data. Third, Web of Science is not complete and may overlook some collaborative publications. Fourth, author name disambiguation was far from perfect, as issues such as spelling variants, name changes, and affiliation changes remained unresolved. Fifth, with small numbers of positive samples, the performance of learned prediction models was not high and needed to be improved. Sixth, disciplinary characteristics which are driving factors of the proposed concepts were not fully discussed due to lack of domain knowledge.

In the following, we discuss our research questions in detail.

RQ1: How do we incorporate co-authorship types and the temporal nature of collaboration into a co-authorship-based framework that can be used to predict long-term research collaboration?

In a coauthored study, different co-authorship types between two researchers describe their individual roles and contributions to the study as well as their mode of collaboration. Collaboration data, accumulated to a certain degree, we then can find patterns of two researchers' collaboration from this history of various combinations of co-authorship types. For a collaborative relationship to be sustained and become a long-term research collaboration, both researchers should realize that each of them has something novel to continuously contribute and that they are complementary to one other. This realization takes time, and the prediction of long-term research collaboration should adequately consider the dynamic within the two researchers' collaboration history. The study introduced eight combinations of co-authorship types and framed the prediction problem as a time series prediction problem that used derived time series features and machine learning.

RQ2: How important are dependency, reciprocity, and informal mentorship in predicting long-term research collaboration?

We found that dependency, reciprocity, and informal mentorship have different roles in different disciplines. Among the three, informal mentorship was important in predicting long-term research collaboration in Agriculture, Geology, and LIS. The predication results in these three disciplines were much higher than in the other two disciplines, and this is partly because informal mentorship, which is a strong type of interpersonal relationship, played a significant role in prediction. Specifically, while the traditional informal

mentorship relationship between junior and senior researchers where senior researchers play a supervising role was important in Agriculture, informal mentorship in both directions was important in Geology and LIS. We see that as collaboration continues over decades in scientific activities, different types of informal mentorship emerge. Reciprocity, which measures interdependence between two researchers, was important to prediction in Agriculture and Geology. These two disciplines had the highest rate of collaboration, where more than 90 % of papers were outputs of collaboration. Considering that these collaboration rates are high, there may be many potential collaborators in these disciplines, and thus reciprocity is an important factor for any two researchers to maintain long-term research collaboration. Finally, dependency was important in all the disciplines to varying degrees of importance. Specifically, only the dependency of less experienced researchers on more experienced ones is important to prediction in Geology and Sociology, whereas in LIS, the opposite holds true. This is partly because LIS is a highly interdisciplinary field where the continuous introduction of new knowledge from young collaborators is necessary to perform high-impact research.

Our experimental results showed that, in predicting long-term research collaboration, the importance of informal mentorship, dependency, and reciprocity had varied across disciplines, possibly due to the inherent differences among the nature of each discipline. In terms of informal mentorship, AlShebli et al. (2020) conducted a survey and discovered that informal mentorship relationships exist ubiquitously in the shared co-authorship. They also found that although a high proportion of mentees reported that they had received suggestions on different research skills from their informal mentors who were listed as their coauthors, the extent of this mentoring varied across different disciplines. This partly explains that the importance of informal mentorship in predicting long-term research collaboration has disciplinary differences. As for dependency, drawing on data from four research fields, Lee et al. (2010) employed a clustering method to identify dependency patterns in research collaboration. The results showed that dependency patterns exist in three research fields, including alternative energy, water shortage, and global warming. This makes it understandable why dependency is important in the disciplines with varying degrees of importance. Due to differences in the research paradigm (Kuhn, 1963), teamwork is more prevalent in Agriculture and Geology than in the other three disciplines we investigated (See Fig. 3, top left). If the scientific output of a discipline is largely composed of papers of single authors, reciprocity would lose its significance in the context of our study. This may explain why reciprocity matters in the fields of Agriculture and Geology. Taken together, abovementioned studies partly support the findings of our study that informal mentorship, dependency, and reciprocity have varying roles in different disciplines in the prediction of long-term research collaboration.

5. Conclusions

In this study, we examined the roles of dependency, reciprocity, and informal mentorship in the prediction of long-term research collaboration in five disciplines. We used co-authorship matrix-based multivariate time series features and interpretable machine learning to train long-term collaboration prediction models and interpret the feature importance of the trained models. We demonstrated the usefulness of the used approaches and explained the different roles of dependency, reciprocity, and informal mentorship in the investigated disciplines. With little existing literature on the prediction of long-term research collaboration, this study provided a conceptual framework of approaching the research problem which can be referenced and refined further by future studies to tackle other important research questions. The introduction of time dimension to the problem and supporting results clearly showed the dynamic nature of research collaboration and the importance of capturing temporal changes for better prediction. While not comprehensive, we proposed three concepts that are associated with long-term research collaboration which may be utilized to formulate and answer fundamental causal questions related to long-term research collaboration. For our future work, we plan to investigate the evolving roles of dependency, reciprocity, and informal mentorship in long-term research collaboration. In addition, future researchers may investigate the association between different long-term co-authorship types and scientific innovation, which is helpful for scholars to promote the efficacy of scientific collaborations and optimize collaboration strategies. This study has some implications for science policymakers. First, because there are differences across disciplines in predictive factors for long-term collaboration, policymakers may consider implementing discipline-dependent strategies of promoting long-term collaborations. For example, in the fields of Agriculture, Geology, and Library and Information Science, supporting informal mentorship programs may facilitate long-term collaborations between mentors and mentees. Second, although long-term collaborations in academia are known to boost researchers' productivity and impact (Petersen, 2015), it is rate. By utilizing the predictive models proposed in the study, governments may establish fundings to encourage collaborative research and support research teams that have high possibilities of becoming long-term research teams.

CRediT authorship contribution statement

Yongjun Zhu: Conceptualization, Data curation, Funding acquisition, Project administration, Writing – original draft. **Donghun Kim:** Methodology, Writing – original draft. **Ting Jiang:** Software, Writing – original draft. **Yi Zhao:** Data curation. **Jiangen He:** Validation. **Xinyi Chen:** Visualization. **Wen Lou:** Methodology, Project administration, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Yonsei University Research Grant of 2022 (2022-22-0394).

References

Abramo, G., D'Angelo, C. A., & Costa, F. D. (2019). The collaboration behavior of top scientists. *Scientometrics, 118*(1), 215–232. https://doi.org/10.1007/s11192-018-2970-9

Abramo, G., D'Angelo, C. A., Costa, F. D., & Solazzi, M. (2011). The role of information asymmetry in the market for university–industry research collaboration. *The Journal of Technology Transfer, 36*(1), 84–100. https://doi.org/10.1007/s10961-009-9131-5

Abramo, G., D'Angelo, C. A., & Murgia, G. (2017). The relationship among research productivity, research collaboration, and their determinants. *Journal of Informetrics, 11*(4), 1016–1030. https://doi.org/10.1016/j.joi.2017.09.007

Agoramoorthy, G. (2017). Multiple first authors as equal contributors: Is it ethical? *Science and Engineering Ethics, 23*(2), 625–627. https://doi.org/10.1007/s11948-016-9794-x

AlShebli, B., Makovi, K., & Rahwan, T. (2020). The association between early career informal mentorship in academic collaborations and junior author performance. *Nat Commun, 11*(1), 5855. https://doi.org/10.1038/s41467-020-19723-8

Amjad, T., Ding, Y., Xu, J., Zhang, C., Daud, A., Tang, J., & Song, M. (2017). Standing on the shoulders of giants. *Journal of Informetrics, 11*(1), 307–323. https://doi.org/10.1016/j.joi.2017.01.004

Bozeman, B., Fay, D., & Slade, C. P. (2013). Research collaboration in universities and academic entrepreneurship: The-state-of-the-art. *The Journal of Technology Transfer, 38*(1), 1–67. https://doi.org/10.1007/s10961-012-9281-8

Bu, Y., Ding, Y., Liang, X., & Murray, D. S. (2018). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology, 69*(3), 438–448. https://doi.org/10.1002/asi.23966

Bu, Y., Ding, Y., Xu, J., Liang, X., Gao, G., & Zhao, Y. (2018). Understanding success through the diversity of collaborators and the milestone of career. *Journal of the Association for Information Science and Technology, 69*(1), 87–97. https://doi.org/10.1002/asi.23911

Bukvova, H. (2010). Studying research collaboration: A literature review. *Sprouts: Working Papers on Information Systems, 10*(3), 1–17.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cheng, M. Y., Hen, K. W., Tan, H. P., & Fok, K. F. (2014). Patterns of co-authorship and research collaboration in Malaysia. *Aslib Proceedings, 65*(6), 659–674. https://doi.org/10.1108/ap-12-2012-0094

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing, 307*, 72–77.

Creamer, E. G. (2004). Assessing outcomes of long-term research collaboration. *Canadian Journal of Higher Education, 34*(1), 27–46.

Cronin, B., Shaw, D., & Barre, K. L. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science & Technology, 54*(9), 855–871.

Cronin, B., Shaw, D., & Barre, K. L. (2004). Visible, less visible, and invisible work: Patterns of collaboration in 20th century chemistry. *Journal of the American Society for Information Science & Technology, 55*(2), 160–168.

Fox, M. F., & Faver, C. A. (1984). Independence and cooperation in research: The motivations and costs of collaboration. *The Journal of Higher Education, 55*(3), 347–359.

Gazni, A., & Thelwall, M. (2014). The long-term influence of collaboration on citation patterns. *Research Evaluation, 23*(3), 261–271. https://doi.org/10.1093/reseval/rvu014

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy, 39*(5), 662–673. https://doi.org/10.1016/j.respol.2010.01.012

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1–18. https://doi.org/10.1016/s0048-7333(96)00917-1

Kuhn, T. S. (1963). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lapidow, A., & Scudder, P. (2019). Shared first authorship. *Journal of the Medical Library Association, 107*(4), 618–620. https://doi.org/10.5195/jmla.2019.700

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology, 66*(7), 1323–1332. https://doi.org/10.1002/asi.23266

Lee, B., Kwon, O., & Kim, H.-J. (2010). Identification of dependency patterns in research collaboration environments through cluster analysis. *Journal of Information Science, 37*(1), 67–85. https://doi.org/10.1177/0165551510392147

Lee, D. H., Seo, I. W., Choe, H. C., & Kim, H. D. (2012). Collaboration network patterns and research performance: The case of Korean public research institutions. *Scientometrics, 91*(3), 925–942. https://doi.org/10.1007/s11192-011-0602-8

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*(5), 673–702. https://doi.org/10.1177/0306312705052359

Levitt, J. M., & Thelwall, M. (2016). Long term productivity and collaboration in information science. *Scientometrics, 108*(3), 1103–1117. https://doi.org/10.1007/s11192-016-2061-8

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy, 29*(1), 31–40. https://doi.org/10.1016/s0048-7333(99)00031-1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res., 12*, 2825–2830.

Petersen, A. M. (2015). Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences, 112*(34), E4671–E4680.

Savanur, K., & Srikanth, R. (2010). Modified collaborative coefficient: A new measure for quantifying the degree of research collaboration. *Scientometrics, 84*(2), 365–371. https://doi.org/10.1007/s11192-009-0100-4

Shen, H., Xie, J., Ao, W., & Cheng, Y. (2022). The continuity and citation impact of scientific collaboration with different gender composition. *Journal of Informetrics, 16*(1), Article 101248. https://doi.org/10.1016/j.joi.2021.101248

Traoré, N., & Landry, R. (1997). On the determinants of scientists' collaboration. *Science Communication, 19*(2), 124–140.

Tsai, C.-H., & Lin, Y.-R. (2016). Tracing and predicting collaboration for junior scholars. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion* (pp. 375–380). https://doi.org/10.1145/2872518.2890516

Wang, W., Liu, J., Yang, Z., Kong, X., & Xia, F. (2019). Sustainable collaborator recommendation based on conference closure. *IEEE Transactions on Computational Social Systems, 6*(2), 311–322. https://doi.org/10.1109/tcss.2019.2898198

Wang, W., Xia, F., Wu, J., Gong, Z., Tong, H., & Davison, B. D. (2021). Scholar2vec: Vector representation of scholars for lifetime collaborator prediction. *ACM Transactions on Knowledge Discovery from Data (TKDD), 15*(3), 1–19. https://doi.org/10.1145/3442199

Wray, K. B. (2006). Scientific authorship in the age of collaborative research. *Studies in History and Philosophy of Science Part A, 37*(3), 505–514. https://doi.org/10.1016/j.shpsa.2005.07.011

Yarime, M., Takeda, Y., & Kajikawa, Y. (2009). Towards institutional analysis of sustainability science: A quantitative examination of the patterns of research collaboration. *Sustainability Science, 5*(1), 115. https://doi.org/10.1007/s11625-009-0090-4

Ye, Q., Li, T., & Law, R. (2011). A coauthorship network analysis of tourism and hospitality research collaboration. *Journal of Hospitality & Tourism Research, 37*(1), 51–76. https://doi.org/10.1177/1096348011425500