IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 September 2022, accepted 20 September 2022, date of publication 26 September 2022, date of current version 7 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210129

RESEARCH ARTICLE

Link Prediction for Statistical Collaboration Networks Incorporating Institutes and Research Interests

XIZHUORAN SONG[1], **YAN ZHANG**[2], **RUI PAN**[1], **AND HANSHENG WANG**[3]

[1]School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China
[2]School of Economics, Xiamen University, Xiamen 361005, China
[3]Guanghua School of Management, Peking University, Beijing 100871, China

Corresponding authors: Rui Pan (panrui_cufe@126.com) and Yan Zhang (zhangyan_elyssa@163.com)

The work of Rui Pan was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 11971504, and in part by the Disciplinary Funding of Central University of Finance and Economics, and in part by the Emerging Interdisciplinary Project of the Central University of Finance and Economics. The work of Hansheng Wang was supported in part by the National Natural Science Foundation of China under Grant 11831008, and in part by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science under Grant KLATASDS-MOE-ECNU-KLATASDS2101.

ABSTRACT An interesting application of the link prediction technique is detecting the potential new links in collaboration networks. In this study, we construct collaboration networks based on the co-authorship information of the papers published in 43 statistical journals from 2001 to 2018. We construct training and testing networks according to the timestamps of the papers and construct a classification dataset for link prediction. We calculate 20 similarity indices based on the training network to perform link prediction. Additionally, we consider nodal attributes (institutes and research interests) to develop two novel predictors called the same institute (SIN) and keywords match count (KMC). Several machine-learning classifiers including support vector machine, XGBoost and random forest are implemented to combine all predictors. After incorporating SIN and KMC, we observe that the area under the receiver operating characteristic curve values of all classifiers improved, indicating that SIN and KMC can significantly improve classification accuracy. Finally, we provide collaborative recommendations for researchers based on the proposed model.

INDEX TERMS Collaboration network, link prediction, nodal attribute, similarity-based approach.

I. INTRODUCTION

Scientific collaboration has the advantages of saving costs and diffusing ideas and insights among collaborators [1]. Therefore, establishing new collaboration links among scientists is one of the main drivers of scientific progress, and it is important to conduct a statistical analysis of scientific collaboration. A scientific collaboration network is a popular tool for analyzing and modeling the relationships among scientific collaborators based on co-authorship [2], [3]. Many studies have been conducted on scientific collaboration networks in different disciplines, including biology [4], [5], physics [6], mathematics [1], computer science [7], and statistics [8]. Through network analysis, interesting results on community development, research patterns, and trends [9] can be extracted.

With an ever-increasing number of researchers, it is not always possible to determine which researchers should collaborate. Therefore, it is important to develop techniques to generate collaborative recommendations. In the context of network science, the collaboration relationships can be described by the collaboration networks, in which nodes represent authors and links between two authors represent they have at least published one paper together. Recommending partnerships is type of a *link prediction* problem [10]. Link prediction is the task of estimating the likelihood of an unobserved link between two nodes [11], [12]. Such estimates are generated based on the information of other observed links and the attributes of nodes. Many studies have been conducted on link prediction in collaboration networks. In the

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

field of theoretical high-energy physics and lattice high-energy physics, Chuan et al. [13] find that the accuracy of link prediction can be improved by considering the similarities among authors, their papers, and common co-authors. In the field of quantum communication, Lande et al. [14] use authors' information and keywords in articles to predict collaboration among scientists. Instead of focusing on a particular discipline, Tuninetti et al. [15] restrict their analysis to papers published in Physical Review Letters. They demonstrate that scientific credit and common scientific interests are predictive of new collaborations among scientists.

Link prediction is a prolific subject in network research. Kumar et al. [16] classify link prediction methods into three categories: similarity-based metrics, probabilistic and maximum likelihood methods, and dimension reduction approaches. Similarity-based metrics are the simplest and most widely used methods. A similarity score is calculated between pairs of nodes, where higher-scoring pairs tend to have more links between them [12], [16]. Similarity-based metrics can be broadly divided into topology-based and node-based metrics [17]. Topology-based metrics are based on topological information such as the number of common neighbors (CN) [18] and parameter dependence [19]. Node-based metrics primarily use the attributes and actions of nodes, which can represent an individual's interests or social behaviors [17]. Attributes can be abstracted from node profiles or metadata. For example, Bhattacharyya et al. [20] study the similarity of users based on keywords from their profiles. Tuninetti et al. [15] use mutual citations and common keywords to predict collaboration between two authors. Both topology-based and node-based metrics provide abundant information for link prediction from different perspectives. Therefore, it is natural to combine these methods to analyze networks.

In this study, we mainly use similarity-based metrics to perform link prediction in collaboration networks. Specifically, we introduce 20 topology-based and two node-based indices. Based on the calculation of these metrics, we explore collaboration recommendations and the following research questions are addressed.

- What are the properties or features of the collaboration network of statisticians, particularly from a dynamic perspective?
- In addition to topology-based metrics, can the similarity metrics among nodes be extracted from other information? Are these metrics helpful for link prediction? Previous research has mainly focused on topology-based metrics, but some studies have shown that node-based metrics are also helpful for link prediction. This study proposes two node-based metrics called the same institute (SIN) and keywords match count (KMC) metrics, which improve predictive performance.
- How can we integrate multiple predictors to recommended collaborators? Previous studies have mainly used a single predictor to recommend collaborators (e.g., [10], [21]). However, different predictors may represent different collaboration characteristics. We apply several machine learning classifiers to combine different predictors for link prediction. Specifically, we combine similarity-based metrics and learning-based frameworks to solve the link prediction problem.

Our analysis makes the following main contributions to the literature. First, we provide a comprehensive comparison of 20 similarity indices, including local, quasi-local, and global indices, for collaboration networks based on statistics. Second, author institutes and research interest are considered in this study and the results prove that these information can improve the performance of link prediction. Third, we apply a link prediction method to statistical papers from 2001 to 2018 to derive useful insights into the comparisons among different disciplines.

The remainder of this paper is organized as follows. Section II explores the characteristics of collaboration networks and presents a dataset for link prediction. Section III describes similarity-based and nodal-attribute-based predictors for link prediction. Section IV presents comparison results between different similarity indices and models. Collaboration recommendations are also discussed in this section. Section V summarizes our conclusions.

II. DATA DESCRIPTION

In this section, we introduce detailed information regarding the dataset developed in this study and the construction of a collaboration network. We also present a statistical analysis of the collaboration network. Furthermore, to perform link prediction, we construct training and testing networks according to the time stamps of papers. Finally, we introduce three types of relationships in dynamic networks and formulate them for data processing for link prediction.

A. THE COLLABORATION NETWORK

Our publication dataset is collected from the "Web of Science Core Collection" (*www.webofscience.com*). Specifically, we first select 43 statistical journals and then collect all of the publications from these journals from January 2001 to May 2018. The journals are listed in Table 8. An example of a publication is presented in Table 1. The paper's title, publication year, keywords, and authorship information are obtained. Based on this information, we can construct the collaboration network with abundant nodal attributes such as an author's institute and research interests. Additionally, collaboration relationships can be studied at the institutional or regional level, which can be explored in future research. After cleaning the data, we identify 47,546 unique authors. Two authors are considered to be co-authors if they have published at least one paper together. To describe the relationships among authors, we construct a collaboration network. In this network, a node represents an author, and an edge (i.e., link) represents a collaborative relationship. The network contains 95,666 edges and the density is only 8.46×10^{-5}, indicating an extremely sparse network. Mathematically, we let

TABLE 1. Example of a paper published in the Annals of Statistics.

Year	Title
2018	Convexified Modularity Maximization for Degree-Corrected Stochastic Block Models
Authorship Information	**Keywords**
Yudong, Chen @Cornell University, USA Xiaodong, Li @University of California Davis, USA Jiaming, Xu @Purdue University, USA	community detection, modularity maximization, degree-corrected stochastic block model, convex relaxation, k-medians, social network

$A = (a_{ij}) \in \mathbb{R}^{n\times n}$ denote the adjacency matrix, where n is the number of authors. If author i collaborates with author j, then $a_{ij} = 1$. Otherwise, $a_{ij} = 0$. We always set $a_{ii} = 0$ for $1 \le i \le n$. It should be noted that the collaboration network is undirected, meaning that $a_{ij} = a_{ji}$.

In a collaboration network, the nodal degree represents the number of unique collaborators for an author, that is, $d_i = \sum_{j\neq i} a_{ij}$. Fig 1 is the histogram of nodal degree in the collaboration network. It can be observed that the distribution is highly right-skewed. After detecting, we find that more than half of the authors have only one or two collaborators in our network, whereas a few authors have a large number of collaborators. Authors with high degrees are identified as important nodes in the network. Specifically, the author with the largest degree (Professor Narayanaswamy Balakrishnan at McMaster University) has 292 collaborators. He has published 385 papers in 43 statistical journals over 18 years. For the node with the largest degree, we extract its second-order neighborhoods and preserve all edges included in these neighborhoods. The resulting network is presented in Fig. 3. The node with the largest degree is in the center and other nodes with high degrees are also labeled in this figure. To further study the characteristics of the structure of the collaboration network, we extract the core network of the entire collaboration network. A q-core network is obtained by removing nodes whose number of neighbors is less than q, as well as the edges connected to them [22]. Fig. 2 shows the sizes of a series of core networks when $q = 1, 2, \ldots, 10$. One can see that the sizes of core networks become stable when $q \ge 6$, which shows the 6-core network removes nodes at the periphery and contains most of the core nodes. Therefore, we choose the 6-core network for further analysis, and it is presented in Fig. 4. A core network is helpful for understanding the most important parts of a network. In the 6-core network, we find that the node in the center is Raymond J Carroll, with a degree of 109. When considering the importance of neighbors, the center node in Fig. 4 also plays a significant role in forging strong links between statisticians.

Transitivity is an important property of social networks [23], [24]. It refers to the tendency of two nodes to form a mutual relationship if they are connected to a

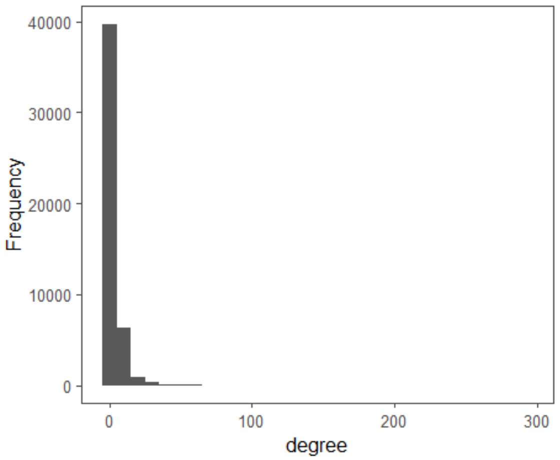


FIGURE 1. The distribution of nodal degrees in the collaboration network. It is highly right-skewed.

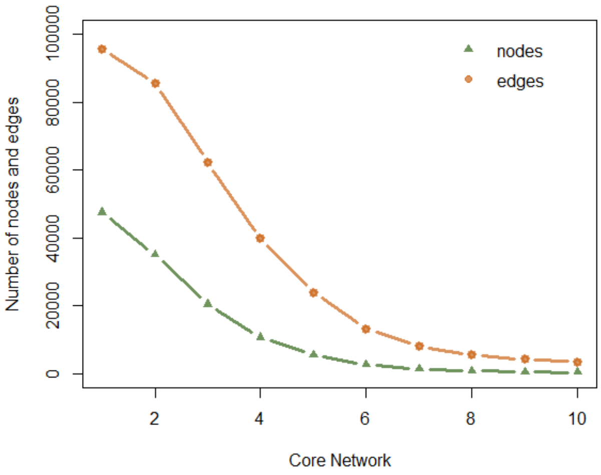


FIGURE 2. The number of nodes and edges in a series of core networks. The x-axis represents different values of q, $q = 1, 2, \ldots, 10$. When q is small, the size of the network changes drastically. When $q \ge 6$, the sizes of core networks become stable.

third node. A network with high transitivity is clustered, and there are communities in which nodes are densely connected. This phenomenon can be observed in Fig. 4, where many groups of researchers are observed to have close collaborative relationships. For example, the group around John Shawe-Taylor includes 58 authors. These authors can be further categorized into three subgroups. We detect that there are two papers written by 38 and 35 authors, which exactly define the left and right subgroups of authors, respectively. The authors who participate in both studies lie in the center subgroup. John Shawe-Taylor, who has links with all the authors in this group, also has links with authors outside the group, indicating that he plays a key role in connecting the group to other authors in the entire network. This also reflects why some of the similarity indices in link prediction are constructed based on common neighbors. Newman [24] has used the clustering coefficient to describe transitivity in collaboration networks, which is defined as $C = 3T/V$,

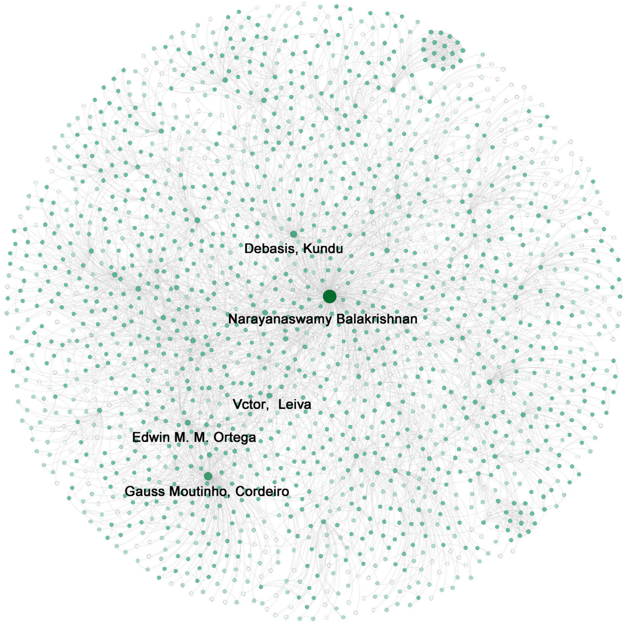


FIGURE 3. Second-order neighborhood network of the node with the largest degree. This network contains 1,745 nodes and 4,148 edges. The density is 0.003. The greater the degree of a node, the darker the color of the node and the greater the size of the node.

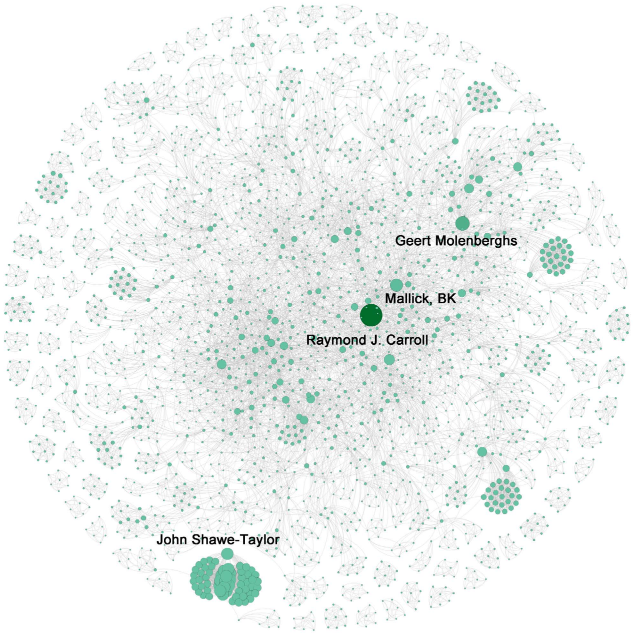
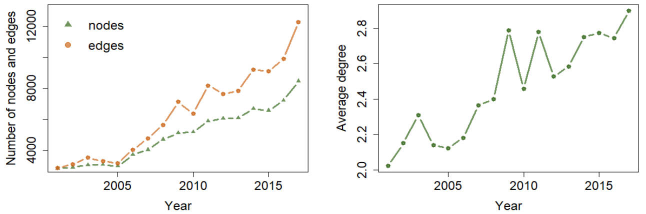


FIGURE 4. The 6-core of the collaboration network. This network contains 2,675 nodes and 13,322 edges. The density is 0.004. The greater the degree of a node, the darker the color of the node and the greater the size of the node. A great number of groups in which nodes are densely connected can be observed. The clustering coefficient of this network is 0.797.

where T denotes the number of triangles in the network and V denotes the number of connected triples. A triangle is a set of three nodes i, j, k satisfying $a_{ij}a_{jk}a_{ik} = 1$, and a connected triple is a single node that is connected to two other nodes. In a collaboration network, the clustering coefficient represents the likelihood of cooperation between two researchers if they have collaborated with a third researcher [24]. In our network, the clustering coefficient is 0.23. Compared to the collaboration networks in other fields, the clustering coefficient of our network is much higher than that of the biological network (0.066). However, it is lower than that in theoretical physics networks, which ranges from 0.33 to 0.43 [24].



(a) Number of nodes and edges (b) Average degree

FIGURE 5. Nodes, edges, and average degrees in each year's network. Both graphs exhibit increasing trends, indicating that researchers have become more collaborative over the past 17 years.

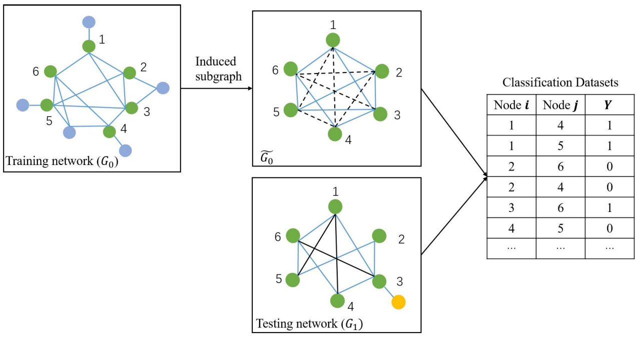
To explore the dynamics of these collaboration patterns, we construct collaboration networks year-by-year from 2001 to 2017. Fig. 5(a) presents the number of nodes and edges (i.e., collaborative relationships) observed over these years, both of which exhibit an increasing trend. This indicates that an increasing number of researchers are collaborating with each other and that new collaborations are emerging. Specifically, the number of researchers in the network increased from 2,811 to 8,466 during this period, and the number of collaboration relationships increased from 2,844 to 12,271. Fig. 5(b) presents the average nodal degrees from 2001 to 2017. The average number of collaborators per author increased from 2.02 to 2.90, indicating that researchers have become more collaborative over the past 17 years. Therefore, it is of great importance to predict and recommend new collaborators to researchers, which can be accomplished through link prediction in collaboration networks.

B. DATASET FOR LINK PREDICTION

Recall that our dataset represents the period from 2001 to 2018. To perform link prediction, the collaboration network is split into two parts according to the timestamp $t = 1, \cdots, 18$. Specifically, we use data from 2001 to 2015 to construct an original collaboration network. As some similarity indices have been designed for connected graphs [10], we consider the giant component $G_0 = (V_0, E_0)$ as a training network, where V_0 is the node set and E_0 is the edge set of G_0. The giant component is the largest connected subgraph in the network. In this case, it contains 72.0% of the nodes (i.e., 26,943) in the original network (2001 to 2015). Motivated by the study of Chuan et al. [13], which leaves 3-years data for testing, we use the data from 2016 to 2018 to construct a testing network $G_1 = (V_1, E_1)$, where V_1 and E_1 are the node and edge sets of G_1, respectively. Additional details regarding the two networks are presented in Table 2.

TABLE 2. Details of the training and testing networks. The time period, number of nodes and edges, density, and average degree are reported. The density of G_1 is slightly higher than that of G_0, whereas the average degree is lower than G_0. This is because G_0 covers a longer timespan.

Network	Time	No. of nodes	No. of edges	Density	Average degree
Training(G_0)	2001-2015	26,943	60,420	0.017%	4.49
Testing(G_1)	2016-2018	16,734	27,973	0.020%	3.34



Node i	Node j	Y
1	4	1
1	5	1
2	6	0
2	4	0
3	6	1
4	5	0
...	...	...

FIGURE 6. Process for constructing the classification dataset. The set of green nodes refers to V. Blue circles and yellow circles represent nodes that only appear in G_0 or G_1, respectively. $\widetilde{G}_0$ is the subgraph derived from G_0 based on the vertex set V. The dotted lines in $\widetilde{G}_0$ refer to nodal pairs that do not have a link. Black lines in G_1 represent newly appearing links among V. For each of the non-linked nodal pairs in $\widetilde{G}_0$, we determine if it has a link in G_1 and accordingly label it as positive or negative. In the classification dataset, $Y = 0$ indicates negative and $Y = 1$ indicates positive.

We now discuss the dynamic patterns of collaborative relationships among authors during these two periods. We let $D_{ij} = (a_{ij}, a_{ji})$ denote the relationship between nodes i and j. There are two types of pairwise relationships in the collaboration network: mutual relationships $D_{ij} = (1, 1)$, and null relationships $D_{ij} = (0, 0)$. Additionally, we let D^0_{ij} and D^1_{ij} denote the relationships between the nodes in G_0 and G_1, respectively. According to Kim and Diesner [25], three types of relationships in dynamic collaboration networks are of particular interest. The first type includes $D^0_{ij} = (1, 1)$ and $D^1_{ij} = (1, 1)$. It represents the collaborative relationships that exist in both the training and testing networks, indicating an ongoing collaborative relationship between authors i and j. The second type includes $D^0_{ij} = (0, 0)$ and $D^1_{ij} = (1, 1)$. This type of relationships involve scenarios in which potential links can be predicted. This type of a phenomenon has attracted significant interest in link prediction problems. The third type is $D^1_{ij} = (1, 1)$, where at least one of the nodes i or j is not in the training network. It is difficult to predict these types of relationships because the training data lack information regarding newly appearing nodes. Motivated by previous studies on link prediction, we mainly focus on the second type of nodal pair in our research.

Our goal is to identify the nodal pairs that belong to the second type of relationship. Whether a new link forms in the future between a pair of nodes can be seen as a dependent variable. From this perspective, link prediction is a binary classification problem. Therefore, it is necessary to construct classification dataset. The process of constructing the classification dataset is illustrated in Fig. 6. First, we obtain the node set $V = V_0 \cap V_1$, which contains nodes in both G_0 and G_1. The size of V is 5,638, which is also the number of unique authors in the classification dataset. These authors are represented by green circles in Fig. 6. The blue and yellow circles represent nodes that only appear in G_0 and G_1, respectively. Then, the subgraph $\widetilde{G}_0$ derived from G_0 based on the node set V is obtained. Next, we select the nodal pairs (i, j) that do not have a link in $\widetilde{G}_0$ (i.e, $i, j \in V$ and $D_{ij} = (0, 0)$). These pairs are represented in Fig. 6 as the dotted lines in $\widetilde{G}_0$. The selected nodal pairs are labeled as positive (1) or negative (0) depending on whether the two nodes have a link in G_1. This classification serves as the dependent variable in the classification dataset. The data are extremely unbalanced. Only a small fraction of nodal pairs generate new links in the testing network. Therefore, we use the undersampling method and randomly remove some negative samples to make the positive samples account for 10% and negative samples account for 90% of the classification dataset [13]. After undersampling, the numbers of positive and negative samples are 3,435 and 30,915, respectively.

III. METHODOLOGY

As mentioned in II-B, link prediction is essentially a binary classification problem. In this section, we illustrate the predictors of the classification problem. Specifically, we introduce the definitions of the 20 similarity indices used in this study. These indices are classified into three categories, as shown in Table 3. We then present a descriptive analysis of the two nodal attributes. We also explain the derivation of novel predictors based on the nodal attributes. The relationships between the novel predictors and dependent variable in the classification dataset are also discussed.

A. SIMILARITY-BASED APPROACHES

First, we introduce the similarity-based indices considered in this study. Similarity-based approaches are widely used for link prediction [16], [17]. For a pair of nodes i and j, a similarity score s_{ij} is calculated. The scores between all pairs of nodes in the network are represented by a symmetric matrix $S = (s_{ij}) \in \mathbb{R}^{n \times n}$. It is observed that nodal pairs with higher scores are more likely to form links.

Table 3 lists the 20 similarity indices considered in our study. According to Kumar et al. [16], these indices can be grouped into three categories of local, quasi-local, and global indices. One of the most commonly used local indices is the CN, proposed by Newman [18]. We use $\Gamma(i)$ to denote the set of neighbors of node i (i.e., $\Gamma(i) = \{i' : a_{ii'} = 1\}$). Furthermore, $|\Gamma(i)|$ denotes the size of $\Gamma(i)$, which is the nodal degree of i and is exactly equal to d_i. Consequently, CN (i, j) is defined as $|\Gamma(i) \cap \Gamma(j)|$, which represents the number of common neighbors of nodes i and j. A series of indices based on CN have been proposed. For example, the Jaccard

(JC) index [26] represents the normalization of CN. Other similarity indices related to CN include the Salton cosine similarity (SC) [27], Sørensen index (SI) [28], Leicht-Holme-Newman (LHN) index [29], hub promoted (HP) index, and hub depressed (HD) index [30]. The preferential attachment (PA) index differs from these indices [31] and is defined as the product of the nodal degrees d_i and d_j; this indicates that the nodes with higher nodal degrees are more likely to form new links. Additionally, it is also natural to consider the degrees of common neighbors, where neighbors with high degrees tend to contribute less to the network. Two examples of this kind of indices are Adamic-Adar (AA) index [32] and resource allocation (RA) index [33]. One can see that most of the local indices are neighbor-based methods, so they require little computational time. In contrast, quasi-local indices require additional computational time because they use more information from the graph. For example, the shortest path (SP) index [10] is the inverse of $d(i, j)$, where $d(i, j)$ is the length of the shortest path between nodes i and j. The local path (LP) index [34] considers the number of paths of lengths two and three between nodes i and j. SP and LP are both path-based methods.

Global indices use the complete topological information of a network [16], leading to higher computational complexity. The Katz index and LHN global (LHNG) index are path-based methods from a global perspective. The Katz index [35] considers all paths between two nodes. We let $|paths_{i,j}^{\langle l\rangle}|$ denote the number of paths of length l ($l \geq 1$) between nodes i and j. The Katz index is the sum of $|paths_{i,j}^{\langle l\rangle}|$ with a weight β^l, where β^l is a predetermined parameter. The LHNG index is a variant of the Katz Index. In the definition of LHNG [29], m is the total number of edges, λ_1 is the greatest eigenvalue in the adjacency matrix A, $D = (d_i) \in \mathbb{R}^{n\times n}$ is a diagonal matrix, and α is a free parameter (see Table 3). Some global indices are related to random walks in a graph and are called random-walk-based methods. In the average commute time (ACT) index, $m(i, j)$ refers to the average number of steps taken by a random walker starting at node i and reaching node j for the first time. The normalized ACT (NACT) introduces $\pi_i = d_i/\sum_i d_i$ as a normalization term [36]. The random walking with restarting (RWR) index also adopts the concept of a random walk. Specifically, in Table 3, q_{ij} is the probability that a random walker starts at node i and is located at node j in the steady state [37]. Other indices include L^+, the cosine based on L^+ (cos L^+), and the matrix forest index (MFI), which are related to the Laplacian matrix $L = D - A$. Specifically, L^+ denotes the pseudoinverse of the Laplacian matrix. According to Fouss et al. [38], each node can be represented by a vector that forms a Euclidean space. The elements of L^+, namely $l_{ij}^+ = [L^+]_{ij}$, represent the inner product between node vectors. Therefore, L^+ can be considered as a similarity matrix. The cosine based on L^+ is the cosine of the angle between the same node vectors as L^+. The MFI differs from the above indices and is based on the matrix forest theorem [39].

TABLE 3. Definitions of 20 similarity indices.

Local indices	
Common Neighbors	$\mathrm{CN}(i,j) = \lvert\Gamma(i) \cap \Gamma(j)\rvert$
Jaccard Index	$\mathrm{JC}(i,j) = \lvert\Gamma(i) \cap \Gamma(j)\rvert / \lvert\Gamma(i) \cup \Gamma(j)\rvert$
Salton Cosine Similarity	$\mathrm{SC}(i,j) = \lvert\Gamma(i) \cap \Gamma(j)\rvert / \sqrt{d_i * d_j}$
Sørensen Index	$\mathrm{SI}(i,j) = 2\lvert\Gamma(i) \cap \Gamma(j)\rvert / (d_i + d_j)$
Hub Promoted Index	$\mathrm{HP}(i,j) = \lvert\Gamma(i) \cap \Gamma(j)\rvert / \min\{d_i, d_j\}$
Hub Depressed Index	$\mathrm{HD}(i,j) = \lvert\Gamma(i) \cap \Gamma(j)\rvert / \max\{d_i, d_j\}$
Leicht-Holme-Newman Index	$\mathrm{LHN}(i,j) = 2\lvert\Gamma(i) \cap \Gamma(j)\rvert / (d_i * d_j)$
Preferential Attachment	$\mathrm{PA}(i,j) = d_i * d_j$
Adamic-Adar Index	$\mathrm{AA}(i,j) = \sum\limits_{z \in \lvert\Gamma(i) \cap \Gamma(j)\rvert} (logd_z)^{-1}$
Resource Allocation Index	$\mathrm{RA}(i,j) = \sum\limits_{z \in \lvert\Gamma(i) \cap \Gamma(j)\rvert} d_z^{-1}$
Quasi-Local indices	
Shortest Paths	$\mathrm{SP}(i,j) = d(i,j)^{-1}$
Local Path Index	$\mathrm{LP} = A^2 + \alpha A^3$
Global indices	
Katz Index	$\mathrm{Katz}(i,j) = \sum\limits_{l=1}^{\infty} \beta^l \lvert paths_{i,j}^{\langle l\rangle}\rvert$
Leicht-Holme-Newman Global Index	$\mathrm{LHNG} = 2m\lambda_1 D^{-1}(I - \frac{\alpha}{\lambda_1}A)^{-1}D^{-1}$
Average Commute Time	$\mathrm{ACT}(i,j) = [m(i,j) + m(j,i)]^{-1}$
Normalized Average Commute Time	$\mathrm{NACT}(i,j) = [m(i,j)\pi_j + m(j,i)\pi_i]^{-1}$
Random Walk with Restart	$\mathrm{RWR}(i,j) = q_{ij} + q_{ji}$
L^+ directly	$\mathrm{S} = L^+$
Cosine based on L^+	$\mathrm{S}(i,j) = l_{ij}^+ / \sqrt{l_{ii}^+ l_{jj}^+}$
Matrix Forest Indexc	$\mathrm{MFI} = (I + L)^{-1}$

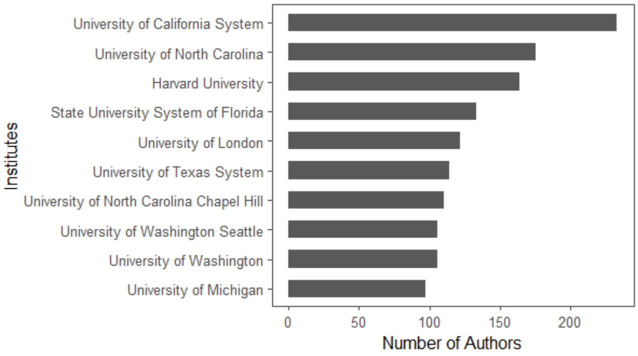


FIGURE 7. Top-10 most-frequent institutes. The University of California System includes UC Berkeley, UCLA, UC Santa Barbara, and many other universities.

B. NODAL-ATTRIBUTE-BASED PREDICTORS

One can see that the similarity indices focus only on the network topology. However, nodal attributes can also provide information for link prediction. We can derive a series of predictors using nodal attributes. Specifically, we take advantage of two attributes in this study, namely, institutes and research interest. For institutes, we let T_i denote the set of institutes of author i, which contains the institutes (referred to as universities in most cases) at which the author worked during the period of 2001 to 2015. Among the 5,638 authors in the classification dataset, 65.9% are affiliated to one or two institutes, indicating that researcher affiliations typically do not change over a short period of time. Fig. 7 presents the 10 most-frequent institutions in the dataset. The University of California System ranks first with 233 authors. Excluding the University of London, all the universities are located in the United States.

Intuitively, if two authors work at the same institute during the same period, they are more likely to collaborate with each other. Some related works [10], [24], [40] have highlighted this concept, but none have applied it to collaboration networks at the author level. Accordingly, we construct a variable called the same institute (SIN) variable. This categorical variable holds true if two authors work at the same institute, is false if they do not work at the same institute, and is unknown if the institute is missing for either of the authors. Therefore, the SIN can be expressed as

$$\mathrm{SIN}(i,j) = \begin{cases} \text{True}, & \text{if } T_i \cap T_j \neq \varnothing; \\ \text{False}, & \text{if } T_i \cap T_j = \varnothing,\ T_i \neq \varnothing \text{ and } T_j \neq \varnothing; \\ \text{Unknown}, & \text{if } T_i = \varnothing \text{ or } T_j = \varnothing. \end{cases} \tag{1}$$

Among the 34,350 pairs of authors, 4% worked at the same institutes during the period from 2001 to 2015, leading to a much higher proportion (74.0%) of collaboration in the future (2016 to 2018). Only 7.3% of the authors whose SIN values are false or unknown later collaborate with each other. It is reasonable for authors working in the same location to collaborate more frequently. Therefore, collaborative relationships can be developed easily [10], [21].

Another nodal attribute is research interest. We extract all keywords from papers written by each author and count the corresponding frequencies. We let W_i denote the keyword set for node i. Fig. 8 presents the word clouds of the top-100 most-frequent keywords. Popular topics include the Markov chain Monte Carlo (MCMC), variable selection, maximum likelihood (ML), EM algorithms, and model selection methods. Table 4 lists the frequencies of the top-20 keywords. According to Wang et al. [41], if the research interests of two authors are similar, they are usually more likely to collaborate. We use the KMC to quantify the similarity between the research interests of two authors. According to Al Hasan et al. [42], this method is simple but effective. Specifically, for nodes i and j, we obtain $W_{ij} = W_i \cap W_j$ as a co-keyword and let $|W_{ij}|$ denote the size of W_{ij}. For each word w_k in W_{ij}, we let n_{ik} and n_{jk} denote the frequencies of word w_k in nodes i and j, respectively. The KMC is defined as follows:

$$\mathrm{KMC}(i,j) = \sum_{k=1}^{|W_{ij}|} (n_{ik} + n_{jk}). \tag{2}$$

A high KMC value indicates that two authors have similar research areas. Therefore, they are more likely to form collaborative relationships. Evidence shows that the authors who share the same keywords tend to collaborate with each other. The proportion of collaboration among such authors is 26.7%, which is much higher than the rate of 6.3% for those who do not share a common research interest. Additionally, only 18.1% of the author pairs share the same keywords and the highest value of the KMC is 268, indicating a high degree of overlap between the research fields of two authors. It is found that both authors are prolific and participate in 144 and 33 papers from 2001 to 2015, respectively. They have published various articles in both Biometrika and the Journal of the American Statistical Association. Frequently occurring keywords in their papers include nonparametric regression, functional data analysis and mixed models.

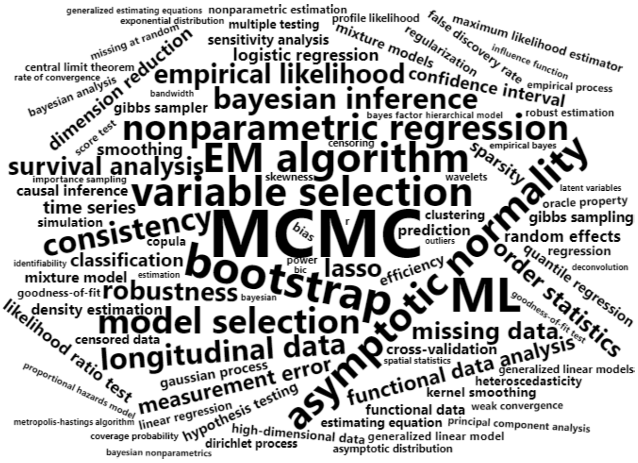


FIGURE 8. Top-100 keywords of papers in 43 statistical journals from 2001 to 2015. Research fields in the MCMC and variable selection methods attract significant interest.

TABLE 4. Top-20 keywords and their corresponding frequencies.

Keywords	Frequency	Keywords	Frequency
MCMC	1361	Bayesian inference	609
ML	977	empirical likelihood	543
bootstrap	918	robustness	537
variable selection	840	survival analysis	517
EM algorithm	837	lasso	506
asymptotic normality	809	missing data	498
nonparametric regression	744	order statistics	476
model selection	710	measurement error	440
consistency	638	functional data analysis	424
longitudinal data	610	classification	377

IV. RESULTS

Link prediction can be performed using any of the aforementioned similarity indices. However, it is also meaningful to use more than one predictor to obtain more accurate results. In this section, we first use the similarity indices separately and then calculate area under the receiver operating characteristic (ROC) curve (AUROC) values. We consider several classifiers, namely a support vector machine (SVM), XGBoost, and random forest (RF), to combine all similarity indices and

TABLE 5. AUROC values of 20 similarity indices.

Similarity indices	AUROC	Similarity indices	AUROC
CN	0.642	SP	0.798
JC	0.642	LP	0.724
SC	0.642	Katz	0.802
SI	0.642	LHNG	0.801
HP	0.642	ACT	0.645
HD	0.642	NACT	0.625
LHN	0.642	**RWR**	**0.829**
PA	0.619	$\boldsymbol{L^+}$	**0.833**
AA	0.642	**cos** $\boldsymbol{L^+}$	**0.833**
RA	0.642	**MFI**	**0.826**

present the results. These classifiers are popularly used in link prediction [43], [44], [45]. To determine the effects of nodal attributes, we incorporate nodal-attribute-based predictors into the models and compare the results. Finally, we present collaborative recommendations for statisticians based on the best model.

A. PERFORMANCE OF SIMILARITY INDICES

The performance of link prediction can be evaluated using ROC curves and corresponding AUROC values [46], [47]. First, we use the 20 similarity indices listed in Table 3 separately to complete link prediction in our statistical collaboration network. Fig. 9 presents the Pearson correlation coefficients of 20 similarity indices [45]. It is clear that most similarity indices have strong positive relationships with each other. Furthermore, there are stronger positive correlations between local and quasi-local indices than between global indices. Indices such as CN and JC, AA and RA, LP and Katz, and L^+ and cos L^+, whose definitions are very similar, exhibit strong correlations. Fig. 10 presents the ROC curves of the indices for the three categories. Generally, global indices outperform quasi-local indices, and quasi-local indices outperform local indices. This is because global indices use the topological information of the entire graph, whereas quasi-local and local indices only use a portion of the information. Among the 10 local indices presented in Fig. 10(a), excluding PA, the prediction ability of the other nine indices is almost the same. This is partly because of the strong correlation among these indices. Fig. 10(b) presents the ROC curves for the SP and LP indices. One can see that SP leads to better performance. The results of the global indices are presented in Fig. 10(c). Katz, LHNG, RWR, L^+, cos L^+, and MFI outperform ACT and NACT. The AUROC values are presented in Table 5. L^+ and cos L^+ exhibit the highest AUROC value of 0.833. RWR ranks second with an AUROC of 0.829, and MFI also performs well. PA exhibits the lowest AUROC value of 0.619, which indicates that using nodal degrees alone to predict new partnerships is very insufficient.

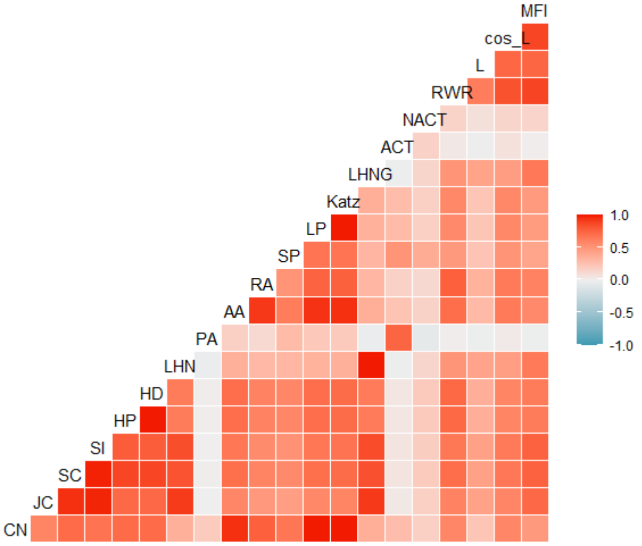
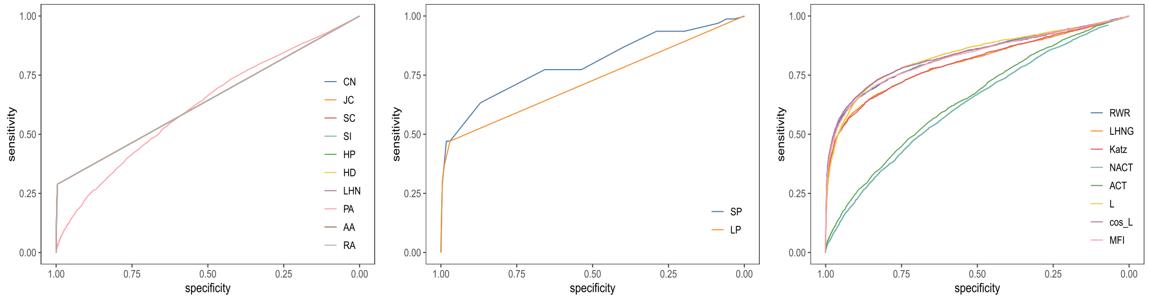


FIGURE 9. Pearson correlation coefficients of 20 similarity indices. Strong positive correlations can be observed.

B. PERFORMANCE OF MACHINE LEARNING APPROACHES

To combine the 20 similarity indices, SIN, and KMC effectively, we apply several machine learning algorithms for classification. Specifically, we use two models and three classifiers to compare the results. The first model contains the 20 similarity indices listed in Table 3 without the nodal-attribute-based predictors, whereas the second model uses all 22 predictors, including KMC and SIN. The machine learning approaches applied in this study are SVM, XGBoost, and RF. These approaches are frequently used in learning-based link prediction methods [43], [44], [45]. An SVM is an extension of support vectors that results from enlarging the feature space using kernels [48]. We apply a linear kernel to our problem. XGboost is a scalable, distributed, and gradient-boosted decision tree machine learning method [49]. RF is an ensemble learning method using decision trees as base learners [50]. The number of trees in all RFs is 500 in this study. In addition to the AUROC, we use the true positive rate (TPR) and precision to evaluate performance. Additionally, the 10 fold cross-validation is implemented for each algorithm.

The performance results are listed in Table 6. One can see that all the three algorithms exhibit improvements when incorporating SIN and KMC, indicating that the nodal-attribute-based predictors can significantly improve prediction accuracy. For example, the value of the AUROC for RF increases from 0.870 to 0.904 when SIN and KMC are added, and the TPR increases from 0.441 to 0.480. The RF with all predictors yields the highest AUROC (0.904). AUROC measures the overall performance of a classification model. An SVM with all predictors yields the highest TPR of 0.483, indicating the ability of SVM to detect newly



(a) Local indices (b) Quasi-local indices (c) Global indices

FIGURE 10. ROC curves of 20 similarity indices in three categories. Global indices outperform the other two categories. L^+ and cos L^+ have the best prediction ability.

TABLE 6. Comparison of models with/without SIN and KMC using SVM, XGBoost, and RF. The RF with both the similarity indices, and SIN and KMC yields the highest AUROC value.

Predictors	Classifiers	AUROC	TPR	Precision
Similarity Indices	SVM	0.809	0.309	0.874
	XGBoost	0.878	0.409	0.800
	RF	0.870	0.441	0.829
Similarity Indices+SIN+KMC	SVM	0.855	0.483	0.778
	XGBoost	0.889	0.442	0.789
	RF	0.904	0.480	0.847

formed links. However, the TPRs of the same predictors in the RF are similar, and the RF yields higher AUROC and precision values. The SVM with only the similarity indices yields the highest precision of 0.874, meaning that it yields the highest proportion of correct predictions among the samples predicted to be positive. However, this model yields the lowest AUROC and TPR values. Considering that finding positive samples is important for solving the link prediction problem, the RF with all predictors outperforms the other combinations.

Next, we explore the effects of each variable. We use two metrics (mean decrease in accuracy and mean decrease in Gini) to measure the importance of variables in the RF (see Fig. 11). One can see that the nodal-attribute-based predictors KMC and SIN are the two most important variables in terms of the metric of mean decrease in accuracy. Furthermore, NACT, RWR, and PA are also important. According to the Gini importance (right side of Fig. 11), the important variables include RWR, cos L^+, SP, MFI, and LHNG.

C. RECOMMENDATIONS

By applying an RF with all the 22 predictors listed in Table 6, we calculate the probability of the collaboration of nodal pairs in our dataset. Collaboration recommendations are based on the results. As mentioned in II-B, the author pairs in our dataset do not collaborate with each other in the training network. There are 40 pairs of authors that are predicted to collaborate, and all of them generate collaborative relationships in the testing network. Table 7 lists these pairs. The last row in this table presents the mean values of the indices. One can see that there are strong similarities between each pair of authors. For example, each pair has at least one common neighbor. Based on the large scale and low density of the training network, 96.7% of the author pairs in the dataset did not have common neighbors. Common neighbors are common collaborators, and researchers can easily learn about each other through common collaborators. The shortest path length between each pair is two, yielding an SP equal to 0.5 for all author pairs. Additionally, the values of L^+ and cos L^+ for these pairs are greater than those of 97% of the author pairs in the dataset. Regarding the values of the nodal-attribute-based predictors, most of the recommended author pairs are related to the same institute and share common keywords, indicating that it is convenient for them to collaborate based on common research interests. Our method can provide useful insights for researchers seeking new potential collaborators. For example, Xueying Zheng and Jie Mao are predicted to collaborate with Guoyou Qin. These three researchers are affiliated to the Fudan University in China and are interested in longitudinal data analyses. Xueying Zheng and Jie Mao published articles on longitudinal data with Guoyou Qin in the Journal of Statistical Computation and Simulation in 2016 and 2018, respectively.

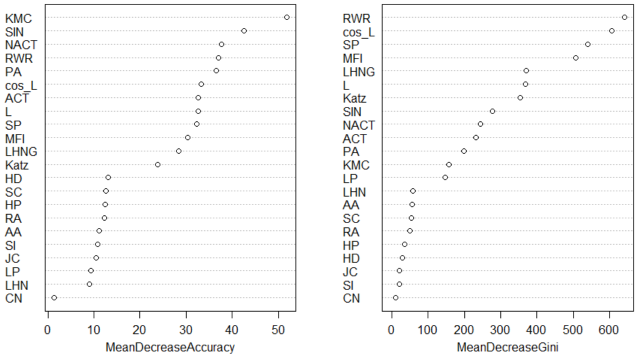


FIGURE 11. Variable importance measured by the mean decrease in accuracy and mean decrease in Gini.

TABLE 7. Top-40 pairs of predicted authors who are most likely to share collaborative relationships. The values of some similarity indices and predictors are also reported for each pair. The mean values for these indices are presented in the last row.

Author i	Author j	CN	L^+	cos L^+	SIN	KMC
Alireza Asgharzadeh	S.M.T.K. MirMostafaee	1	0.02	0.08	True	4
Anthony J. Hayter	Fang Wan	3	0.04	0.17	True	12
Brian Claggett	Marc Alan Pfeffer	3	0.02	0.09	True	0
Brian Williams	Brian Phillip Weaver	1	0.04	0.13	Unknown	6
Changchun Xie	Wanrong Liu	1	0.03	0.09	False	0
Changjun Yu	Xijun Liu	1	0.23	0.35	Unknown	2
Edgar Brunner	Frank Konietschke	4	0.05	0.27	False	6
Guoyou Qin	Xueying Zheng	2	0.02	0.06	True	22
Guoyou Qin	Jie Mao	2	0.02	0.06	True	6
Hardegen, A.	Gopalan Nair	1	0.06	0.16	True	2
Hussein R. Al-Khalidi	Jie Li	1	0.32	0.4	False	0
Italia De Feis	Luisa Cutillo	1	0.04	0.09	Unknown	6
Jan De Neve	Donald John Best	1	0.22	0.33	True	4
Jan De Neve	John Rayner	1	0.19	0.35	True	4
Joanne Wendelberger	Christine Anderson-Cook	2	0.04	0.2	Unknown	2
Julia Braun	Stefanie Muff	1	0.03	0.07	True	0
Limin Peng	Michele Marcus	2	0.03	0.14	True	2
Lisa Doove	Tom Wilderjans	1	0.09	0.17	True	0
Luciana Dalla Valle	Fabrizio Leisen	1	0.04	0.11	False	2
Mahbubul Majumder	Xiaoyue Cheng	1	0.08	0.16	True	2
Miguel Angel Uribe Opazo	Audrey H.M.A. Cysneiros	2	0.03	0.1	True	16
Minh Tang	Yongjin Park	1	0.21	0.36	True	0
Minh-Ngoc Tran	Mattias Villani	2	0.03	0.11	False	12
Muhammad Azam	Mohammad Aslam	1	0.11	0.27	Unknown	0
Noel Veraverbeke	Candida Geerdens	1	0.02	0.06	True	4
Paul Garthwaite	Andre Charlett	1	0.1	0.27	Unknown	2
Pauliina Ilmonen	Klaus Nordhausen	2	0.03	0.13	True	10
R K Milne	Gopalan Nair	1	0.06	0.16	True	2
Seiya Imoto	Hidetoshi Matsui	1	0.43	0.49	True	4
Tom Loeys	Beatrijs Moerkerke	1	0.06	0.1	True	2
Weihua Zhao	Jianbo Li	2	0.03	0.14	True	8
Woncheol Jang	Sunghoon Kwon	1	0.04	0.06	False	0
Woo-Dong Lee	Yongku Kim	1	0.08	0.15	Unknown	0
Xin Chen	Zhihua Su	1	0.04	0.06	False	0
Yan-Yong Zhao	XingFang Huang	1	0.04	0.1	True	0
Yee Whye The	Prunster Igor	2	0.02	0.14	False	20
Yongku Kim	Sang-Gil Kang	1	0.08	0.15	Unknown	0
Zhonghua Li	Qin Zhou	1	0.05	0.17	False	0
Zhonghua Li	Bin Chen	1	0.05	0.15	False	0
Ziding Feng	Adi Gazdar	1	0.02	0.06	True	0
Mean		0.043	0.002	0.004	–	1.32

TABLE 8. List of 43 journals in statistics. The journals are ordered alphabetically by name.

Journals
ADVANCES IN DATA ANALYSIS AND CLASSIFICATION
AMERICAN STATISTICIAN
ANNALS OF APPLIED STATISTICS
ANNALS OF STATISTICS
ANNALS OF THE INSTITUTE OF STATISTICAL MATHEMATICS
ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION
BAYESIAN ANALYSIS
BERNOULLI
BIOMETRICS
BIOMETRIKA
BIOSTATISTICS
COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION
COMMUNICATIONS IN STATISTICS-THEORY AND METHODS
COMPUTATIONAL STATISTICS
COMPUTATIONAL STATISTICS & DATA ANALYSIS
ELECTRONIC JOURNAL OF STATISTICS
INTERNATIONAL STATISTICAL REVIEW
JOURNAL OF APPLIED STATISTICS
JOURNAL OF BUSINESS & ECONOMIC STATISTICS
JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
JOURNAL OF MULTIVARIATE ANALYSIS
JOURNAL OF NONPARAMETRIC STATISTICS
JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
JOURNAL OF STATISTICAL PLANNING AND INFERENCE
JOURNAL OF STATISTICAL SOFTWARE
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES A-STATISTICS IN SOCIETY
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES D-THE STATISTICIAN
JOURNAL OF TIME SERIES ANALYSIS
R JOURNAL
SCANDINAVIAN JOURNAL OF STATISTICS
SPATIAL STATISTICS
STATISTICA SINICA
STATISTICAL ANALYSIS AND DATA MINING
STATISTICAL METHODOLOGY
STATISTICAL METHODS AND APPLICATIONS
STATISTICAL MODELLING
STATISTICAL SCIENCE
STATISTICS
STATISTICS & PROBABILITY LETTERS
STATISTICS AND COMPUTING

V. CONCLUSION

In this study, we collect co-authorship information from 43 statistical journals from 2001 to 2018. We then construct collaboration networks for different time periods based on this information. Statistical analyses are conducted to explore the characteristics and evolution of the collaborative network. The results suggest that researchers have become more collaborative over the past 17 years. We use data from 2001 to 2015 to construct a collaboration network and extract its giant component as a training network. A corresponding testing network is constructed using data from 2016 to 2018. The goal of link prediction is to use information from the training network to predict the possible new links in the testing network. Therefore, a classification dataset is constructed. To predict new links, 20 similarity indices are calculated and two novel nodal-attribute-based predictors are developed. After comparing the prediction capabilities of the 20 similarity indices, we discover that the global indices L^+ and cos L^+ outperform the others. We further improve prediction accuracy by applying machine learning approaches to combine similarity indices with KMC and SIN. The AUROC value of the best model is 0.904, indicating excellent performance. Finally, we apply the best model to the nodal pairs in the dataset and present the top-40 recommended author pairs. They formed collaborative relationships in the testing network, demonstrating that our model has a good recommendation capabilities. However, there are still some limitations of this study which need further advancement. For example, calculating global indices on large-scale networks requires high computational time. It is inefficient so new methods applicable to large-scale networks deserve to be explored.

Several directions for future research are possible. First, additional link prediction approaches can be applied to collaboration networks using statistics. For example, the probabilistic and maximum-likelihood models can be implemented in such networks. The comparison of these approaches is worth studying. Second, collaboration networks can be constructed with different structures such as weighted or directed networks. This could provide more information and lead to improved results. Additionally, our methods can be applied to the networks in various fields such as economics. We can then compare the results of networks across disciplines to draw rich conclusions.

APPENDIX A JOURNAL LIST

See Table 8.

REFERENCES

[1] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5200–5205, Apr. 2004.

[2] M. Coccia and L. Wang, "Evolution and convergence of the patterns of international scientific collaboration," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 8, pp. 2057–2061, Feb. 2016.

[3] M.-G. Hâncean, M. Perc, and J. Lerner, "The coauthorship networks of the most productive European researchers," *Scientometrics*, vol. 126, no. 1, pp. 201–224, Jan. 2021.

[4] M. Tomassini and L. Luthi, "Empirical analysis of the evolution of a scientific collaboration network," *Phys. A, Stat. Mech. Appl.*, vol. 385, no. 2, pp. 750–764, Nov. 2007.

[5] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Apr. 2002.

[6] C. K. Singh and S. Jolad, "Structure and evolution of Indian physics co-authorship networks," *Scientometrics*, vol. 118, no. 2, pp. 385–406, Jan. 2019.

[7] A. Nunes da Silva, M. M. Breve, J. P. Mena-Chalco, and F. M. Lopes, "Correction: Analysis of co-authorship networks among Brazilian graduate programs in computer science," *PLoS ONE*, vol. 17, no. 7, Jul. 2022, Art. no. e0271937.

[8] P. Ji and J. Jin, "Coauthorship and citation networks for statisticians," *Ann. Appl. Statist.*, vol. 10, no. 4, pp. 1779–1812, Dec. 2016.

[9] P. Ji, J. Jin, Z. T. Ke, and W. Li, "Co-citation and co-authorship networks of statisticians," *J. Bus. Econ. Statist.*, vol. 40, no. 2, pp. 469–485, Apr. 2022.

[10] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2003.

[11] N. Barbieri, F. Bonchi, and G. Manco, "Who to follow and why: Link prediction with explanations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1266–1275.

[12] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Statist. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011.

[13] P. M. Chuan, L. H. Son, M. Ali, T. D. Khang, L. T. Huong, and N. Dey, "Link prediction in co-authorship networks based on hybrid content similarity metric," *Int. J. Speech Technol.*, vol. 48, no. 8, pp. 2470–2486, Aug. 2018.

[14] D. Lande, M. Fu, W. Guo, I. Balagura, I. Gorbov, and H. Yang, "Link prediction of scientific collaboration networks based on information retrieval," *World Wide Web*, vol. 23, no. 4, pp. 2239–2257, Mar. 2020.

[15] M. Tuninetti, A. Aleta, D. Paolotti, Y. Moreno, and M. Starnini, "Prediction of new scientific collaborations through multiplex networks," *EPJ Data Sci.*, vol. 10, no. 1, p. 25, Dec. 2021.

[16] A. Kumar, S. S. Singh, K. Singh, and B. Biswas, "Link prediction techniques, applications, and performance: A survey," *Phys. A, Stat. Mech. Appl.*, vol. 553, Sep. 2020, Art. no. 124289.

[17] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: The state-of-the-art," *Sci. China Inf. Sci.*, vol. 58, no. 1, pp. 1–38, Jan. 2015.

[18] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, 2001, Art. no. 025102.

[19] Y.-X. Zhu, L. Lü, Q.-M. Zhang, and T. Zhou, "Uncovering missing links with cold ends," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 22, pp. 5769–5778, Nov. 2012.

[20] P. Bhattacharyya, A. Garg, and S. F. Wu, "Analysis of user keyword similarity in online social networks," *Social Netw. Anal. Mining*, vol. 1, no. 3, pp. 143–158, Jul. 2011.

[21] E. Yan and R. Guns, "Predicting and recommending collaborations: An author-, institution-, and country-level analysis," *J. Informetrics*, vol. 8, no. 2, pp. 295–309, Apr. 2014.

[22] F. W. Crawford, "Discussion of 'coauthorship and citation networks for statisticians,'" *Ann. Appl. Statist.*, vol. 10, no. 4, pp. 1827–1834, Dec. 2016.

[23] C. Aggarwal and K. Subbian, "Evolutionary network analysis: A survey," *ACM Comput. Surv.*, vol. 47, no. 1, pp. 1–36, Jul. 2014.

[24] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, Jan. 2001.

[25] J. Kim and J. Diesner, "Formational bounds of link prediction in collaboration networks," *Scientometrics*, vol. 119, no. 2, pp. 687–706, Mar. 2019.

[26] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull. Soc. Vaudoise Sci. Nat.*, vol. 37, pp. 547–579, Jan. 1901.

[27] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[28] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Biologiske Skrifter*, vol. 5, pp. 1–34, Jun. 1948.

[29] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 73, no. 2, Feb. 2006, Art. no. 026120.

[30] E. Ravasz, A. L. Somera, and D. A. Mongru, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002.

[31] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Phys. A, Statist. Mech. Appl.*, vol. 311, nos. 3–4, pp. 590–614, Aug. 2002.

[32] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Soc. Netw.*, vol. 25, no. 3, pp. 211–230, Jul. 2003.

[33] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, Oct. 2009.

[34] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 4, Oct. 2009, Art. no. 046122.

[35] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, Mar. 1953.

[36] W. Liu and L. Lü, "Link prediction based on local random walk," *Proc. EPL*, vol. 89, no. 5, p. 58007, Mar. 2010.

[37] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Dec. 2006, pp. 613–622.

[38] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.

[39] P. Chebotarev and E. Shamis, "The matrix-forest theorem and measuring relations in small social groups," *Automat. Remote Control*, vol. 58, no. 9, pp. 1505–1514, 1997.

[40] R. Guns and R. Rousseau, "Recommending research collaborations using link prediction and random forest classifiers," *Scientometrics*, vol. 101, no. 2, pp. 1461–1473, Nov. 2014.

[41] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 322–331.

[42] M. A. Hasan, V. Chaoj, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proc. Workshop Link Anal., Counter-Terrorism Secur.*, 2006, vol. 30, no. 9, pp. 798–805.

[43] E. Bütün, M. Kaya, and R. Alhajj, "Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks," *Inf. Sci.*, vols. 463–464, pp. 152–165, Oct. 2018.

[44] Y. Zhu, D. Huang, W. Xu, and B. Zhang, "Link prediction combining network structure and topic distribution in large-scale directed network," *J. Organizational Comput. Electron. Commerce*, vol. 30, no. 2, pp. 169–185, Mar. 2020.

[45] A. Vital and D. R. Amancio, "A comparative analysis of local similarity metrics and machine learning approaches: Application to link prediction in author citation networks," *Scientometrics*, vol. 127, no. 10, pp. 1–18, Aug. 2022.

[46] H. Yuliansyah, Z. A. Othman, and A. A. Bakar, "Taxonomy of link prediction for social network analysis: A review," *IEEE Access*, vol. 8, pp. 183470–183487, 2020.

[47] B. Chen, Y. Hua, Y. Yuan, and Y. Jin, "Link prediction on directed networks based on AUC optimization," *IEEE Access*, vol. 6, pp. 28122–28136, 2018.

[48] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2013.

[49] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[50] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

XIZHUORAN SONG is currently pursuing the B.S. degree with the School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China.

YAN ZHANG received the B.S. degree in mathematics and applied mathematics from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019, and the M.S. degree in applied statistics from the Central University of Finance and Economics, Beijing, in 2021. She is currently pursuing the Ph.D. degree in mathematical statistics with Xiamen University, Fujian, China.

RUI PAN received the Ph.D. degree from Peking University, China, in 2014. She is currently an Associate Professor with the School of Statistics and Mathematics, Central University of Finance and Economics, China. Her research interests include data science, network data analysis, and data mining.

HANSHENG WANG received the Ph.D. degree from the University of Wisconsin–Madison, in 2001. He is currently a Professor with Peking University. He has published more than 50 research articles. His research interests include high-dimensional data analysis, search engine marketing, social network analysis, and deep learning. He is the Elected American Statistical Association Fellow.