Dynamic network analytics for recommending scientific collaborators

Lu Huang[1] • Xiang Chen[1]✉ • Yi Zhang[2] • Yihe Zhu[3] • Suyi Li[1] • Xingxing Ni[1]

[1] School of Management and Economics, Beijing Institute of Technology, Beijing, 100081, P.R. China

[2] Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia

[3] New H3C TECHNOLOGY CO LTD, Hangzhou, 310000, P.R. China

Lu Huang

Affiliation: School of Management and Economics, Beijing Institute of Technology

Address: 5 South Zhongguancun Street, Haidian District, Beijing, P.R. China. Zip code: 100081

E-mail: huanglu628@163.com

Xiang Chen (corresponding author)

Affiliation: School of Management and Economics, Beijing Institute of Technology

Address: 5 South Zhongguancun Street, Haidian District, Beijing, P.R. China. Zip code: 100081

E-mail: bjchenxiang@hotmail.com

Yi Zhang

Affiliation: Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney

Address: 15 Broadway, Ultimo, NSW, Australia. Zip code: 2007

E-mail: yi.zhang@uts.edu.au

Yihe Zhu

Affiliation: New H3C TECHNOLOGY CO LTD

Address: 466 Changhe Road, Binjiang District, Hangzhou, P.R. China. Zip code:310000

E-mail: zhuyihe_bit@163.com

Suyi Li

Affiliation: School of Management and Economics, Beijing Institute of Technology

Address: 5 South Zhongguancun Street, Haidian District, Beijing, P.R. China. Zipcode: 100081

E-mail: lisuyi@bit.edu.cn

Xingxing Ni

Affiliation: School of Management and Economics, Beijing Institute of Technology

Address: 5 South Zhongguancun Street, Haidian District, Beijing, P.R. China. Zipcode: 100081

E-mail: nixingxing11@163.com

✉ Xiang Chen

bjchenxiang@hotmail.com

Abstract:

Collaboration is one of the most important contributors to scientific advancement and a crucial aspect of an academic's career. However, the explosion in academic publications has, for some time, been making it more challenging to find suitable research partners. Recommendation approaches to help academics find potential collaborators are not new. However, the existing methods operate on static data, which can render many suggestions less useful or out of date. The approach presented in this paper simulates a dynamic network from static data to gain further insights into the changing research interests, activities and co-authorships of scholars in a field – all insights that can improve the quality of the recommendations produced. Following a detailed explanation of the entire framework, from data collection through to recommendation modelling, we provide a case study on the field of information science to demonstrate the reliability of the proposed method, and the results provide empirical insights to support decision-making in related stakeholders - e.g., scientific funding agencies, research institutions and individual researchers in the field.

Keywords network analytics • collaboration recommendation • link prediction • semantic analysis • dynamic network

Introduction

Scientific collaboration is essential for enhancing the productivity of researchers (Wang et al. 2019b; Landa et al. 2020) and driving innovation (Lee et al. 2011; Zhang et al. 2017a). The process of selecting collaborators for research is complex, and although affected by multiple factors, e.g., geographical location (Terveen and McDonald 2005) and social context (Lee and Schleyer 2010), the relevance of their research content is crucial as well (Chaiwanarom and Lursinsap 2015). Nowadays, the information overload due to the emergence of big scholarly data is making this process even more time-consuming (Schreiber et al. 2017; Zhou et al. 2017), and it is impossible to identify potential research partners by relying only on personal networks and social relationships. Hence, an effective algorithm for recommending possible collaborators could be useful. Existing methods for recommending scientific collaborators can be broadly categorized into three types: collaborative filtering (CF) (Haruna et al. 2017), content-based methods (Sugiyama et al. 2010), and social network-based methods (Liben-Nowell and Kleinberg 2007; Shibata et al. 2011). Hybrid models of the above strategies have also been proposed (Chen et al. 2018; Işık et al. 2018). However, the following gaps remain in literature: (1) Most previous studies are based on static data and fail to characterize the changes in academic activities, since collaborative relationships change over time (Bian et al. 2014; Zhang et al. 2017b); (2) Data sparsity is a major challenge for conventional recommendation systems, as it may result in information loss and performance reduction (Wang et al. 2019a; Zhang et al. 2021); (3) Conventional text analytics for research content typically focuses on the collinear relationship of keyword frequency while ignoring the internal semantics between keywords (Sun et al. 2018; Zhou and Jiang 2020), which may lead to an inaccurate similarity analysis.

Our study was aimed at developing a novel recommendation methodology for academic collaborations based on dynamic network analytics, considering the research content in scholarly papers as well as the topological structure of bibliometric networks. In other words, the proposed framework considers the changing landscape of academic collaborations and research interests over time. The bibliometric input to the model is simulated as a data stream. The integral tools in the framework include link prediction, text analysis, and machine learning techniques. Specific functions include: (a) link

prediction combined with an autoregressive integrated moving average (ARIMA) model (Box and Jenkins 1976) to transform static co-author network indicators into dynamic ones for analysis; (b) the use of time-sliced author–keyword matrices to track changes in researchers' interests over time and to alleviate data sparsity (Lai and Tsang 2016); (c) the use of a Word2Vec model (Mikolov et al. 2013) to detect underlying semantics in large-scale texts by mapping words from vocabularies to vectors and help further similarity measures; and (d) the use of bagging support vector machine (BSVM) (Kim et al. 2002b) to identify existing collaboration patterns and, from these patterns, recommend new ones.

To demonstrate the reliability and feasibility of our method, we conducted a case study in the information science discipline and compared the results with those obtained using four conventional link prediction baselines and three state-of-the-art baselines. Expert knowledge-based evaluations were performed, and the results not only demonstrate the ability of our method in recommending scientific collaborators, but also shed new light on this field of study and its practices.

Related Work

A brief review of the relevant literature in each research stream follows, including scientific collaboration recommendation, link prediction, and similarity measurements of research interests.

Scientific Collaboration Recommendation

Scientific collaboration is important in scientific research (Sun et al. 2019; Wang et al. 2019b). Through scientific collaboration, researchers from different academic backgrounds having different knowledge skills and thinking modes can form interdisciplinary research teams, thus complementing each other and sharing knowledge, abilities, and resources (Gui et al. 2019). Collaboration networks represent an important medium to examine scholarly communication (Hoang et al. 2019; Kim and Diesner 2019; He et al. 2021).

Scientific collaborators can be recommended using three major methods (Zhang et al. 2019b; Molaei et al. 2021): 1) CF, which makes use of past academic information of researchers, such as historical publications or partnerships, to predict future potential collaborators (Zhang et al. 2019b; Molaei et al. 2021); this method can be further divided into model-based CF (Guo et al. 2016) and memory-based CF (Meyffert et al. 2013); 2) Content-based methods, which recommend collaborators based on content-related features in published literature (Zhang et al. 2019b; Yang et al. 2020), namely titles, abstracts, and keywords (Alinani et al. 2018; Wang et al. 2018), by employing latent Dirichlet allocation (Blei et al. 2003) and its variant as representative models (Li et al. 2018b; Gao et al. 2018); 3) Social network-based methods, which analyze the cooperation mechanism by extracting the characteristics of collaboration networks (Wang et al. 2019b; Yang et al. 2020). In addition, some hybrid methods have been proposed to leverage the specific advantages of the above models and improve the overall recommendation result (Liu et al. 2018; Zhang et al. 2019a; Xia et al. 2021).

Link Prediction

Link prediction means estimating the possibility of either an unknown link or a future link between two nodes in a network, such as a co-author or citation network, that are not connected or do not appear to be connected (Lu and Chen 2020; Chen et al. 2021). There are three main link prediction methods: similarity-based algorithms, maximum likelihood methods, and probabilistic models (Lü and Zhou 2011; Wang et al. 2017).

Link prediction has been widely used in the analysis of scientific collaboration networks (Cho and

Yu 2018; Wang et al. 2019b). Liben-Nowell and Kleinberg (2007) were among the first bibliometricians to study link prediction by graphing co-authorship networks. Since then, many improvements have been made to this technique. For example, its accuracy has been improved by combining multiple static predictors (e.g., common neighbors, Jaccard, and Adamic-Adar (AA)) with machine learning (Benchettara et al. 2010; Guns and Rousseau 2014). Some researchers have predicted links in heterogeneous networks (Li et al. 2018a; Lande et al. 2020). For example, Yang et al. (2020) proposed a link prediction model to dynamically reflect the changes in academic cooperation by analyzing institution networks and co-author networks.

Recently, some novel link prediction methods have been proposed, including the Node2Vec-based approach (Grover and Leskovec 2016), variational graph autoencoder (VGAE) (Kipf and Welling 2016), and stacking models, for nearly optimal link prediction (OLP) (Ghasemian et al. 2019).

Similarity Measurements of Research Interests

Pearson's correlation, Salton's cosine, and Euclidean distance are all common similarity measures and are widely-used in co-citation analysis, co-word analysis, co-authorship analysis (Leydesdorff 2005; Eck and Waltman 2008). For example, Zhao and Strotmann (2010) introduced author bibliographic coupling analysis as a method to map the research activities of active authors for a more realistic picture of the current state of research in a field. However, the purpose of these methods is mainly to establish hidden and indirect academic relationships by means of third-party literature.

Semantic similarity measurements operate on the content of a researcher's work. Divided into two types, knowledge-based and corpus-based (Zhu and Lan 2014), both quantify the degree to which two words are semantically related using information drawn from a source database (Mcinnes and Pedersen 2013). In the case of knowledge-based methods, for example, Meng et al. (2013) used WordNet to measure semantic similarity but with a hybrid measure that combines path, information content, and feature-based approaches. Corpus-based methods have emerged since WordNet stopped being supported and can draw on a variety of different sources (Pradhan et al. 2020). For example, Deerwester et al. (2010) proposed latent semantic analysis as a way to improve the detection of relevant documents based on the terms found in queries. Kong et al. (2016) proposed a novel collaboration recommendation model that uses Word2Vec to generate more accurate feature descriptions of each researcher.

Methodology

Our framework, depicted in Fig. 1, identifies potential collaborations based on two elements: 1) the topological structure of an evolving network; and 2) the attributes of the nodes, which represent the research interests of authors.

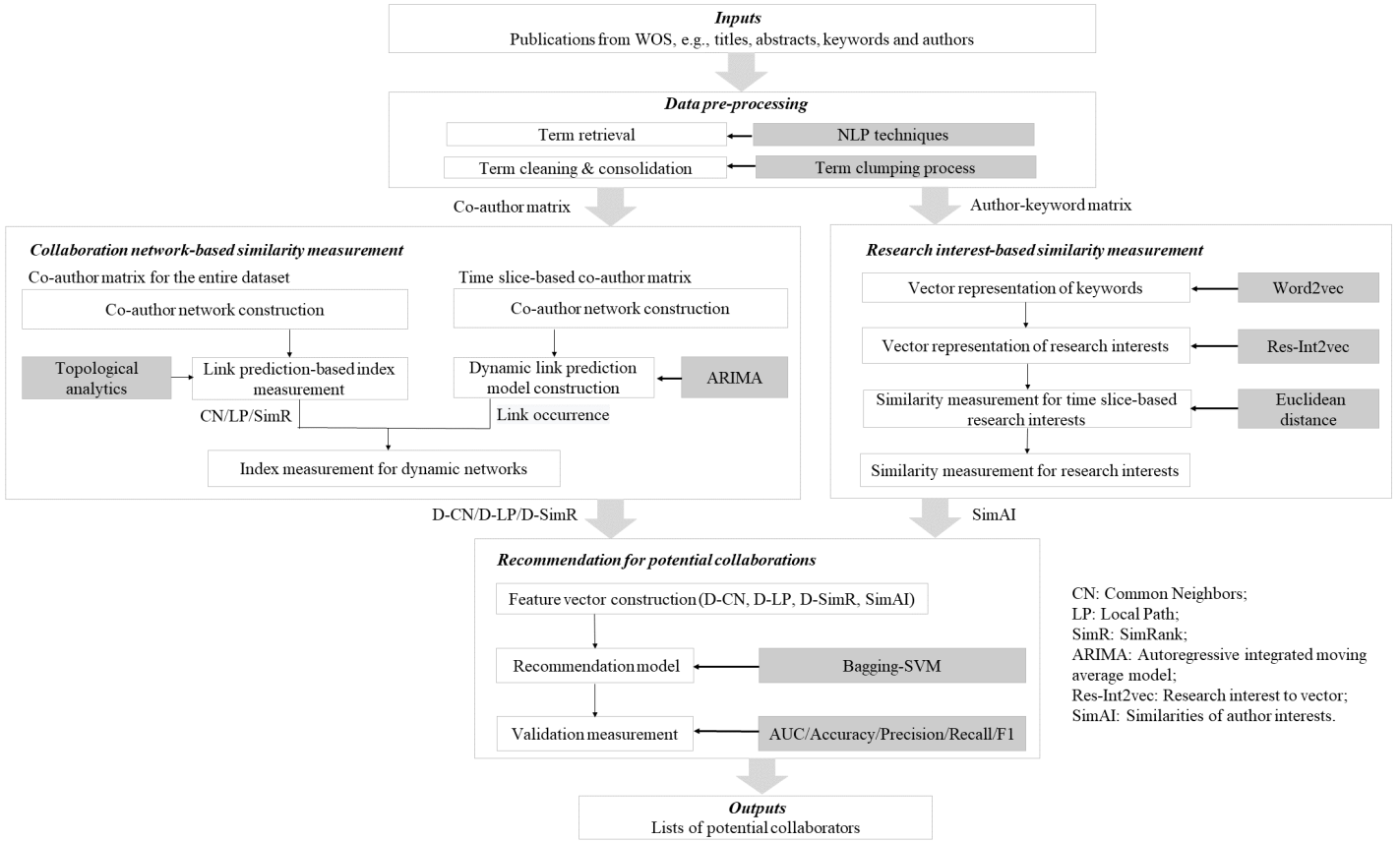


Fig. 1. Our dynamic network analysis framework

Inputs and Data Pre-processing

The input is research articles gathered from the Web of Science (WoS), which include data such as titles, abstracts, keywords and authors. A natural language processing (NLP) technique is used to retrieve terms from the titles and abstracts, and then a term clumping process removes noise, consolidates terms, and identifies core terms (Zhang et al. 2014).

To capture the data's dynamic features, the entire dataset is cut into several slices to simulate a data stream, i.e., based on the publication year of research articles, the entire dataset can be expressed as $\{D_t, t = 1,2, \dots, T\}$, where T denotes the number of slices, and D_t denotes the t^{th} slice.

Collaboration Network-based Similarity Measurement

Co-author Network Construction

The first step in measuring the similarity between authors is to construct two types of co-author networks. One is a static co-author network based on a co-author matrix of the entire dataset. The other is a dynamic network based on the co-author matrices from each time slice. The relationship between the dynamic network G_t and the static network G_s can be described as:

$$G_s = \sum G_t \tag{1}$$

Link Prediction-based Index Measurement

The next step is to calculate three topological indicators of the static network G_s. These are link predictors based on the structural attributes of the network.

(1) Common Neighbors (CN; Newman 2001). Two nodes x and y are likely to have a link if they have sufficient common neighbors. The index of CN S_{xy}^{CN} is calculated as

$$S_{xy}^{CN} = \sum_{Z \in \Gamma(x) \cap \Gamma(y)} w_{xz} + w_{yz} \tag{2}$$

where w_{xz} and w_{yz} denote the weights of the edges between node z and x, and y and z, respectively, $\Gamma_{(x)}$ and $\Gamma_{(y)}$ denote the sets of neighbors of x and y, and $\Gamma_{(x)} \cap \Gamma_{(y)}$ denote the common neighborhoods of x and y.

(2) Local Path (LP; Zhou et al. 2009). This is a path-based similarity index. It considers the contribution of third-order neighbors (the nodes that can be reached by the target node through two edges) based on the common neighbor index. In this paper, this index is adopted for a weighted network. The path distance of the node is defined as the reciprocal of the sum of the reciprocal weights. According to the index, a greater weight of the edge between nodes indicates the less connectivity of the path. Thus, the higher the index, the closer the relationship between nodes. LP S_{xy}^{LP} is calculated as

$$S_{xy}^{LP} = (A^2)_{xy} + \alpha(A^3)_{xy} \tag{3}$$

where $\alpha(0 \le \alpha \le 1)$ is a tunable parameter, A is the adjacency matrix of the network G_s, and $(A^n)_{xy}$ represents the number of paths with a length equal to n.

(3) SimRank (SimR; Jeh and Widom 2002). This indicator reflects the similarity between two nodes on the assumption that the more similar the two nodes are, the more likely they are to be connected in the future. SimR S_{xy}^{SimR} is calculated as

$$S_{xy}^{SimR} = \begin{cases} 1, & x = y \\ \beta \frac{\sum_{v \in \Gamma(x)} \sum_{v' \in \Gamma(y)} S_{vv'}^{SimR}}{n_x n_y}, & x \neq y \end{cases} \tag{4}$$

where n_x and n_y represent the degrees of node x and y (the degree of a node is determined by the

number of its connected edges). Node v is a node in the set of x's neighbors, node v' is a node in the set of y's neighbors, $S_{vv'}^{SimR}$ is the index of SimR between node v and v', and $\beta(0 \leq \beta \leq 1)$ is a decay factor.

Dynamic Link Prediction Model Construction

The above indicators are good measures of the similarities in static networks. However, they tend to profile static topological structures and fail to reflect the varying characteristics over time; this reduces the scientificity and accuracy of the recommendation results. For example, suppose that in a dynamic co-author network with five snapshots over time, author A collaborated with B for 10, 0, 0, 0, 0 times in order, which means that they did not collaborate at all in the subsequent four time slices. We can infer that one of them may have changed his/her research or has retired, in which case it becomes inappropriate to recommend A to B or B to A. However, in the entire static collaboration network, the total frequency of collaboration between A and B is 10, which is high. If the prediction was made directly based on the static network, A was likely to be recommended to B.

The ARIMA model, introduced by Box & Jenkins (1976), is a method for time series analysis. It can predict the data of the $(T+1)^{th}$ time through the time series of the $1^{st} \sim T^{th}$ time. It is represented as (p, d, q), where p is the number of autoregressive integrate terms, q is the number of moving average terms, and d is the number of orders required to convert a nonstationary sequence into a stationary sequence.

We incorporate the ARIMA model into our framework as a method to calculate the weight of a potential link in $(T+1)^{th}$ time based on the time-series constructed by the occurrence frequency of each link in $1^{st} \sim T^{th}$ time, in order to form a dynamic link prediction model. According to Huang and Lin (2009), the optimal parameters of the ARIMA model may fall within $p = 0, 1, 2, 3$; $d = 0, 1$; and $q = 0, 1, 2, 3$. The Akaike information criterion (AIC) can be used as a measure (Akaike, 1974) to assess the quality of each combination and select the model with the lowest AIC score.

The link occurrence score SC_{xy} for the link (x, y) is set to be the potential for the link occurrence frequency in the $(T+1)^{th}$ time, which is greater than 1:

$$SC_{xy} = \Pr(\hat{x}_{xyT+1} > 1) \tag{5}$$

where $\hat{x}_{xyT+1}$ denotes the link occurrence frequency of nodes x and y at the time slice $T+1$, and $\Pr(\hat{x}_{xyT+1} > 1)$ denotes the probability that $\hat{x}_{xyT+1}$ is greater than 1.

Index Measurement for Dynamic Networks

In this part, three static link prediction indices, namely the common neighbors, local path (LP), and SimRank (SimR), are transformed into dynamic indices based on SC_{xy}. Based on the design described by Huang and Lin (2009), three new dynamic indices are introduced. $D - CN_{xy}$ represents the common neighbors $D - LP_{xy}$ represents the local path, and $D - SimR_{xy}$ is the SimRank between the nodes x and y in the dynamic network, calculated as

$$D = \left(S + \frac{M}{\varepsilon}\right) * \left(SC_{xy} + \frac{min_{SC}}{\varepsilon}\right), \varepsilon > 1 \tag{6}$$

where D represents
$D - CN_{xy}$, $D - LP_{xy}$, or $D - SimR_{xy}$, S can be either S_{xy}^{CN}, S_{xy}^{LP}, or S_{xy}^{SimR}, M can represent min_{cn}, min_{LP}, and min_{SimR}, which are the minimum values of S_{xy}^{CN}, S_{xy}^{LP}, and S_{xy}^{SimR} in the static network G_s, min_{sc} is the minimum value in SC_{xy}, and ε is a tunable parameter. From Eq. (6), it is apparent

that the dynamic index is equal to the static index multiplied by the probability generated from the ARIMA model, which aims to realize the combination of static link prediction and time series analysis.

Research Interest-based Similarity Measurement

We chose the Word2Vec model (Mikolov et al. 2013) to represent researchers' interests in the form of vectors, and the Euclidean distance was used to measure the similarity between the interests.

Vector Representation of Keywords

We selected the skip-gram model in the Word2Vec suite to construct keyword vectors (Mikolov et al. 2013). This is because, between the two common modules in Word2Vec—skip-gram and CBOW—the former has proven to have a small advantage with bibliometric data (Zhang et al. 2018; Hu et al. 2018). The inputs for the training are word sequences generated from the text in the abstracts and titles, and the keywords are mapped as a vector originating from a point in a multidimensional semantic space using the skip-gram model.

Vector Representation of Research Interests

The authors' interests are converted into vector representations by loading the keyword vectors created in the previous step into an author–keyword matrix M_k^t. The research interest vector Vri_a^t of the author a in the time slice t can be calculated from

$$Vri_a^t = \sum F_t(k,a) * V_k^t \quad (7)$$

where $F_t(k,a)$ represents the co-occurrence frequency of keyword k and author a in the time slice t, and V_k^t denotes the vector of keyword k in time slice t.

Similarity Measurement for Time Slice-based Author Interests

The similarity between the vectors representing author interests is calculated in each time slice based on the Euclidean distance. In the Word2Vec model, word vectors are normalized, and in this case, we need a more discriminative method to represent the distance between vectors. Therefore, we focus on the absolute distance between the word vectors rather than the relative distance (e.g., cosine measurement) and then transform the similarity in the research content into the Euclidean distance of the research content vector. $Simlar_{ab}^t$ denotes the similarity in the research interests between authors a and b in the time slice t, that is, the similarity between Vri_a^t and Vri_b^t, and it is defined as follows:

$$Simlar_{ab}^t = \frac{1}{\rho_{ab}^t} \quad (8)$$

where ρ_{ab}^t is the Euclidean distance between authors a and b in the time slice t.

Similarity Measurement for Research Interests

Once the similarities for each time slice have been calculated, we determine the overall similarity in the interests between two researchers by adding the vectors. $SimAI_{ab}$ represents the similarity between the research interests of authors a and b, which can be calculated as:

$$SimAI_{ab} = \sum_{t=1}^{T} \gamma^{T-t} Simlar_{ab}^t (0 < \gamma < 1) \quad (9)$$

where $Simlar_{a,b}^t$ denotes the similarity in the research interests between authors a and b in time t, and γ^{T-t} represents the weight of the time slices.

Recommendations for Potential Collaborations

Each of the indicators described above becomes an input vector to the SVM model for classification, and the known collaborations between researchers form the labeled training set. We selected SVM because it includes an approximation algorithm that reduces the computational complexity in terms of both time and dimensionality. Although a single SVM cannot learn the exact parameters of global optimization and therefore cannot provide optimal classification performance on all test examples, according to Kim et al. (2002a), a bagging algorithm can help mitigate this problem. Following Kim et al. (2002b), we therefore applied the BSVM to improve the stability and detection accuracy of the classifier.

Feature Vector Construction

The feature vectors for each author comprise topological indicators, i.e., $(D-CN,\ D-LP,\ D-SimR)$ and the author interest vectors $(SimAI)$.

Recommendation Model

The feature vectors are used as inputs to the model, and any existing collaborations between authors are assembled into a training set for the BSVM algorithm. Existing collaborations are flagged as a binary "author-pair" label, where 1 implies that the two researchers have co-authored a paper, and 0 implies otherwise. The data from time slice 1 to $T-1$ are used for training, and the data in time slice T are used for validation. The final output is a list of potential collaborators for each author.

Validation Measurement

For our case study, we chose five evaluation metrics to verify the accuracy of the recommendation results: Area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F1 (Shani et al. 2001). Our calculation methods for each indicator are explained as follows.

We consider two networks G_{t_1} and G_{t_2} $(t_1 < t_2)$ with a given set of the same continuous time slices. To evaluate a prediction, we calculate the score of each pair of nonexistent nodes in G_{t_1} and then observe whether these pairs of nodes generate edges in G_{t_2}. If the score is higher than or equal to a pre-set threshold, a connection between the two nodes is very likely in the future. If it falls below the threshold, a future connection is unlikely. Thus, with these statistical predictions, it is possible to build a confusion matrix, such as the one presented in Table 1 (Junuthula et al. 2016).

Table 1. Link prediction confusion matrix

		Prediction	
		Existing Edges	Nonexistent Edges
Actuality	Existing Edges	True Positive, TP	False Negative, FN
	Nonexistent Edges	False Positive, FP	True Negative, TN

(1) AUC. The value of the AUC is the area under the ROC curve, which can be derived by sweeping the threshold using the false positive rate (FPR) as the abscissa and the true positive rate (TPR) as the ordinate. The formulae for the two rates are as follows:

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{TP+TN} \tag{10}$$

If all the scores are generated from an independent and identical distribution, the AUC value should be approximately 0.5. Therefore, a value exceeding 0.5 reflects the extent to which the algorithm outperforms pure chance (Lü and Zhou 2011). As per the recommendation made by Fawcett (2005) and

Huang et al. (2018), we set the threshold as 0.5, because our goal was to identify all the factors conducive to the collaboration as much as possible.

The formulae for the other evaluation metrics are as follows:

(2) Accuracy:

$$Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)} \tag{11}$$

(3) Precision:

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

(4) Recall:

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

(5) F1:

$$F1 = \frac{2(Precision*Recall)}{Precision+Recall} \tag{14}$$

Empirical Study

We chose Information Science (IS) as the research field to conduct an in-depth analysis to demonstrate the application of our method. IS is a relatively mature field, with large amount of literature data and more common scientific collaboration. Furthermore, as information scientists ourselves, it is convenient to find experts who are willing to help us verify the reliability of the recommendation results. However, this is not to say that our methodology is only suitable for this research area; it can be extended to broader research fields.

Following the studies conducted by White and McCain (2010) and Hou et al. (2018), we selected 22 leading IS journals. As listed in Table 2, 12 of the journals were the highest-ranked from 2005 to 2008; the remaining 10 were the highest-ranked from 2009 to 2016.

Table 2. Sources of journal data

No	**Journal Name (2005–2008)**	**Journal Name (2009–2016)**
1	Annual Review of Information Science and Technology	Scientometrics
2	Information Processing & Management (and Information Storage & Retrieval)	Journal of the Association for Information Science and Technology
3	Journal of the American Society for Information Science[1]	Information Research an International Electronic Journal
4	Journal of Documentation	Journal of Informetrics
5	Journal of Information Science	Information Processing and Management
6	Library & Information Science Research (and Library Research)	Journal of Documentation
7	Proceedings of the American Society for Information Science (and electronic or digital libraries;	Journal of Information Science

[1] JASIST changed its name from *Journal of the American Society for Information Science and Technology* to *Journal of the Association for Information Science and Technology* in 2014.

	Proceedings of the ASIS Annual Meeting)	
8	Scientometrics	Library and Information Science Research
9	Electronic Library	Research Evaluation
10	Information Technology and Libraries (and Journal of Library Automation)	ARIST
11	Library Resources & Technical Services	
12	Program—Automated Library and Information Systems	

Inputs and Data Pre-processing

Using the Web of Science (WoS) database, we retrieved a total of 4,593 articles publishing in these journals between 2005 and 2016, from which VantagePoint's[2] natural language processing function helped extract 5,068 authors and 86,718 terms. After term clumping to remove noise and consolidate synonyms/same authors, 777 authors and 4,758 terms remained. More data on this process are provided in Tables 3 and 4.

Table 3. Stepwise results of the term clumping process for author names

Step	Description	Terms
0	raw author name	5068
1	name disambiguation	4421
2	remove authors less than three time steps	777

Table 4. Stepwise results of the term clumping process for keywords

Step	Description	Terms
0	raw terms retrieved by NLP technique	86718
1	remove single words	74815
2	remove terms starting/ending with nonalphabetic characters, e.g., "10%"	71442
3	normalize to British spelling	69714
4	remove common terms in scientific article, e.g., "introduction"	55182
5	consolidate terms based on expert knowledge, e.g., "information retrieval" and "IR"	47513
6	consolidate terms with the same stem	31742
7	remove meaningless terms	28030
8	remove words less than 3	4758

Table 5 lists the descriptive statistics of the co-author network (2005 to 2016). Note that the weight of the nodes denotes the number of articles associated with this node, and the weight of the edges represents the co-occurrent frequency between its connected nodes.

Table 5. Descriptive statistics of co-author network (2005 to 2016)

	Weight			
	Max.	***Min.***	***Mean***	***Standard deviation***
Nodes	153	1	8.0077	14.065
Edges	72	1	2.7458	4.0172

[2] https://www.thevantagepoint.com/

Considering the stability of academic research and the granularity of data, the entire dataset was divided into six two-year slices (see Table 6), and co-author and author–keyword matrices were constructed for each time slice based on the co-occurrence data.

Table 6. Detailed information on the six sequences

Time Slice	2005–2016	T1: 2005–2006	T2: 2007–2008	T3: 2009–2010	T4: 2011–2012	T5: 2013–2014	T6: 2015–2016
Nodes	777	437	547	621	575	555	214
Edges	1049	235	318	480	388	364	234
Density	0.0038	0.0025	0.0021	0.0025	0.0024	0.0024	0.01

Similarities in the Collaboration Network

With the matrices built, we constructed a full co-author network covering the entire period (2005 to 2016) using VOSviewer (Eck and Waltman 2010), as shown in Fig. 2, and calculated the three link predictors, i.e., CN using Eq. (2), LP using Eq. (3), and SimR using Eq. (4).

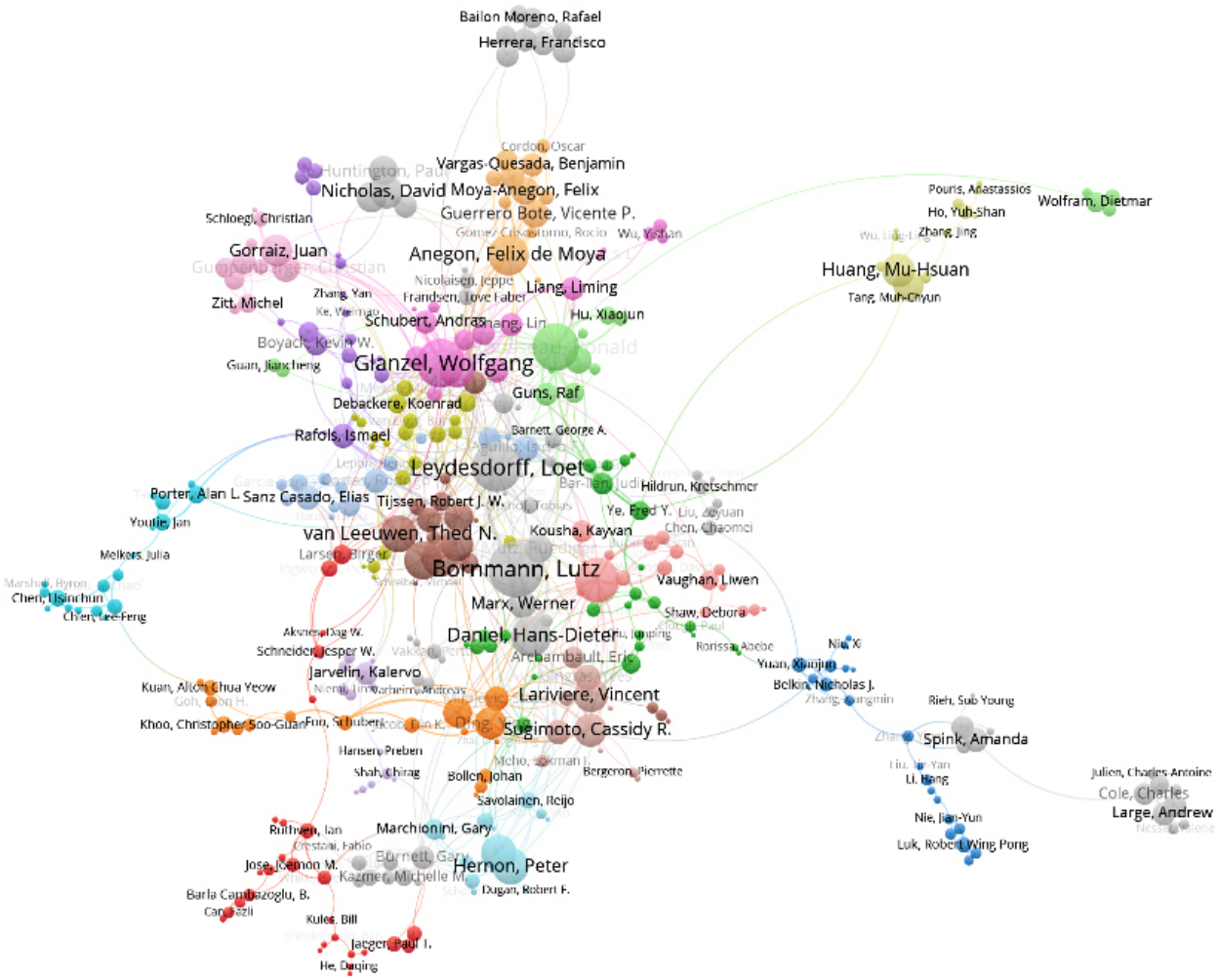
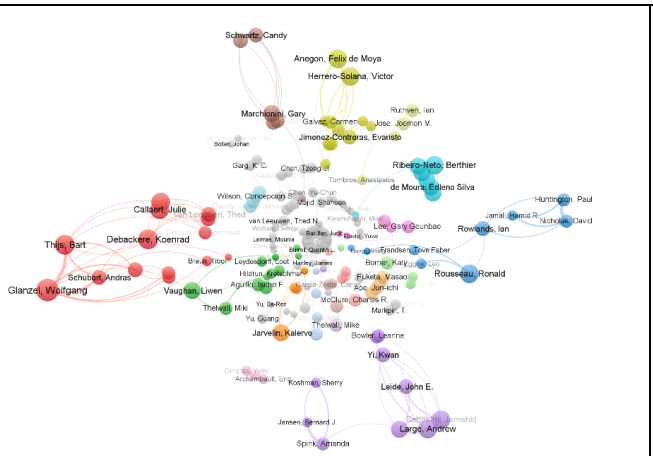


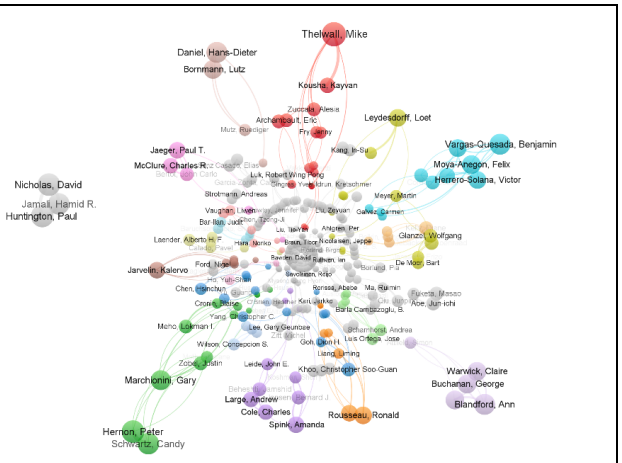
Fig. 2. Co-author network (2005 to 2016)

We also generated a co-occurrence network of the authors for each of the six time slices. These appear

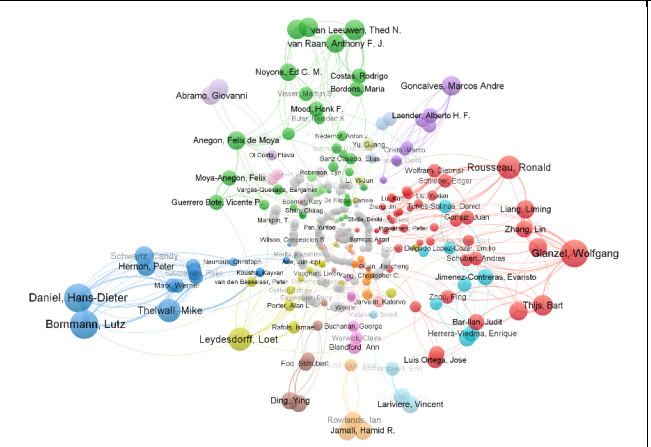
in Fig. 3. These networks show changes in the authors' relationships over time. For example, some collaboration teams have been relatively stable for a long duration, such as the ones indicated by the red and purple areas in Fig. 3(a), whereas the others are far more dynamic. For example, the collaboration between Bornmann and Leydesdorff was not apparent until 2011 (red nodes in Fig. 3(d)) and, since then, many other collaborators have co-authored with these researchers in various combinations.



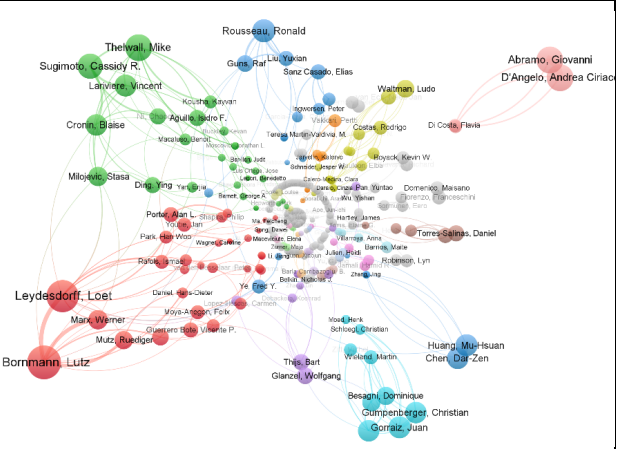
(a)



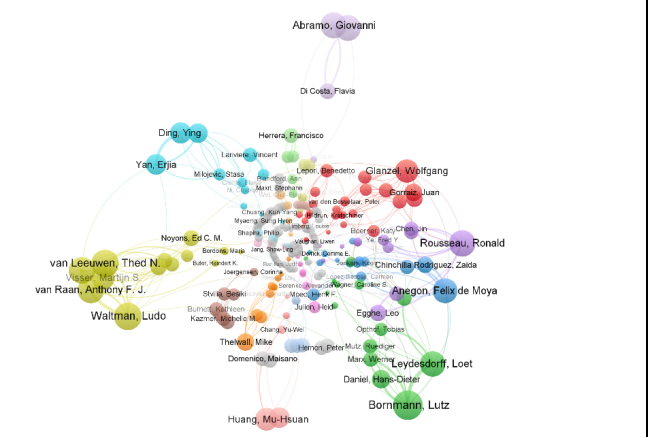
(b)



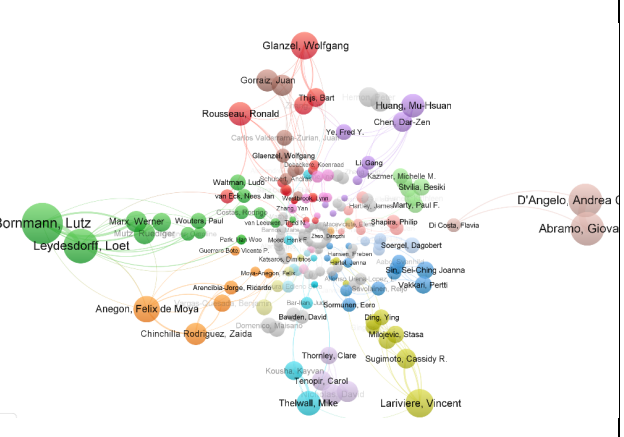
(c)



(d)



(e)



(f)

Fig. 3. Co-author networks by time slice: (a) from 2005 to 2006, (b) from 2007 to 2008, (c) from 2009 to 2010, (d) from 2011 to 2012, (e) from 2013 to 2014, and (f) from 2015 to 2016.

Similarities Between Researchers' Interests

The next stage was to calculate the similarities between researchers' interests. We trained the skip-gram model based on the results of the term clumping with the Gensim package in Python. Since higher dimensions have been shown to capture better semantics (Wang et al. 2015), we set the number of dimensions of the vectors to 300, and the keywords were converted into semantic-level vectors using the trained model.

Subsequently, based on the author–keyword matrix M_k^t and the keyword vectors, we constructed the overall author interest vectors (Eq. (7)), for each time slice (Eq. (8)), and for each author pair *SimAI* (Eq. (9)).

With a total of 2,120 author pairs, we calculated the dynamic link predictors $SimAI$, $D-LP$, $D-CN$, and $D-SimR$. Table 7 provides the descriptive statistics.

Table 7. Descriptive statistics for the author pairs

	Maximum value	Minimum value	Average value	Standard deviation
$\boldsymbol{SimAI}$	0.9688	0.0000	0.4226	0.2406
$\boldsymbol{D-LP}$	0.0513	0.0000	0.0026	0.0035
$\boldsymbol{D-CN}$	5.8698	0.0000	1.1706	1.0803
$\boldsymbol{D-SimR}$	1.4627	0.0003	0.0194	0.0760

Recommendations for Potential Collaborators

Our method described above suggests using the data from time slice 1 to time slice $T-1$ as the training set (approximately 70% of the dataset), with the remaining 30% T as the validation set. However, when we tried this split, we found that there were few partnerships in T. Additionally, compared with the other time slices, there were significantly more changes in this time slice. Therefore, to ensure data coherence, we used the time slice 1 to time slice $T-2$ as the training set (70%) and $T-1$ as the validation set (30%).

Evaluation with Statistical Metrics

We selected five standard quantitative metrics (AUC, accuracy, precision, recall, and F1) as our validation measurements.

First, we conducted a pre-training on the dataset of 1000 and adjusted the parameters γ with a step size of 0.1 (with $\alpha = 0$ and $\beta = 0$). Following the experiments conducted by Huang and Lin (2009), we set ε in Eq. (6) as 2. The results showed a relatively stable AUC value when γ = [0.1–0.4], [0.5–0.8], and [0.9]. Therefore, we selected the first value of each interval and set γ to 0.1, 0.5, and 0.9, representing low importance, intermediate importance, and high importance, respectively. Subsequently, we conducted a pre-training on the dataset of 1000 (with $\gamma = 0.1$) and adjusted the parameters α and β with a step size of 0.1. The results showed that our method exhibited the best performance in terms of the AUC when $\alpha = 0.2$ and $\beta = 0.1$. Therefore, we set $\alpha = 0.2$ and $\beta = 0.1$.

Second, we compared our method with four conventional static indicators (common neighbors, AA, PA, and Jaccard). The calculation method of the common neighbors is expressed in Eq. (2), and the methods of the other indicators are given below:

(1) Preferential attachment (PA; Zhou et al. 2009). This index considers the information of neighboring nodes and has been proven to be the best neighbor-based predictor for collaborations (Yan and Guns 2014). PA S_{xy}^{PA} can be expressed as

$$S_{xy}^{PA} = n_x n_y \tag{15}$$

where n_x and n_y denote the degrees of the nodes *x* and *y*, respectively.

(2) Adamic-Adar (AA; Adamic and Adar 2003). This index assigns more weight to neighbors with fewer connections, rather than simply counting the common neighbors (Adamic and Adar 2003). AA S_{xy}^{AA} can be expressed as

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log n_z} \tag{16}$$

where n_z represents the degree of the node z.

(3) Jaccard (Liben-Nowell and Kleinberg 2007). This index can be considered as the normalization of the common neighbors (Guns and Rousseau 2014). Jaccard $S_{xy}^{Jaccard}$ is calculated as

$$S_{xy}^{Jaccard} = \frac{\sum_{z \in \Gamma(x) \cap \Gamma(y)} w_{xz} + w_{yz}}{\sum_{z \in \Gamma(x) \cup \Gamma(y)} w_{xz} + w_{yz}} \tag{17}$$

where $\Gamma(x) \cup \Gamma(y)$ is the set of neighbors of *x* and *y*.

Furthermore, we compared our methods with three state-of-the-art link prediction methods, including Node2Vec-based approach, VGAE, and stacking models, for nearly OLP.

Tables 8 and 9 list the experimental results, from which we can make the following observations:

(1) When γ = 0.1, our method achieved the best performance. Specifically, we compared the values of each metric under the same data scale when γ took different values and increased the maximum value of each index in Table 8. Regardless of the dataset, our method achieved the best results in terms of AUC and F1 when γ = 0.1. When γ = 0.5, our method produced the best accuracy on the dataset of 1000 and the best recall on the datasets of 500 and 100. When γ = 0.9, our method achieved the best precision on the dataset of 1000.

(2) Compared with the results of the other four static network analysis methods, our method (with γ = 0.1) achieved the best performance in terms of the AUC, accuracy, precision, and F1, regardless of the dataset. On average (Table 9), the AUC value for our method was 8% higher than that for the common neighbors, 19% higher than that for the AA, and 25% higher than those for PA and Jaccard. Meanwhile, the average value of the precision for our method was almost 100% higher than that for the other four static network analysis methods. Note that when the data scale was 1000, both PA and Jaccard achieved a recall of 0.96, outperforming our method. Intuitively, the static similarity indices, such as the common neighbors, PA, AA, and Jaccard, fully capture historical collaborations. In other words, author pairs that have collaborated historically tend to have a high static similarity, leading to a high recall for static network analysis methods when the recommendation set is relatively high but the positive sample size of the test set is relatively small. As mentioned in the methodology section, these static network analyses fail to reflect the characteristics that vary over time, thus resulting in a lower precision, accuracy, and F1.

(3) Compared with the three state-of-the-art methods, our method (with $\gamma = 0.1$) achieved the best performance on all five indicators, regardless of the scale of the dataset (1000, 500, or 100). Specifically, the average value of the AUC for our method was 27% higher than that for Node2Vec and 19% higher than that for VGAE. Furthermore, the average accuracy, average precision, and average F1 for our method were 31%, 37%, and 25% higher than those for OLP, respectively.

From the observations above, we find that the proposed model significantly outperformed the four conventional approaches and the three state-of-the-art link prediction methods.

Table 8. Model performance and method comparison

	Methods	Scale of dataset	Evaluating Indicators				
			AUC	Accuracy	Precision	Recall	F1
Dynamic network analysis	$\gamma = 0.1$	1000	***0.97***	0.92	0.81	**0.92**	***0.86***
		500	***0.91***	***0.87***	***0.87***	0.90	***0.88***
		100	***0.93***	***0.90***	***1.00***	0.91	***0.96***
	$\gamma = 0.5$	1000	0.91	***0.94***	0.73	0.70	0.71
		500	0.90	0.81	0.72	***0.95***	0.82
		100	0.92	***0.90***	0.88	***0.93***	0.91
	$\gamma = 0.9$	1000	0.95	0.84	***0.86***	0.69	0.76
		500	0.89	0.85	0.86	0.84	0.85
		100	0.91	0.89	***1.00***	0.88	0.93
Static network analysis	Common neighbors	1000	0.94	0.86	0.19	0.78	0.31
		500	0.86	0.83	0.39	0.78	0.52
		100	0.77	0.70	0.73	0.78	0.75
	AA	1000	0.77	0.35	0.34	0.92	0.50
		500	0.81	0.41	0.25	0.29	0.27
		100	0.74	0.61	0.26	0.04	0.07
	PA	1000	0.62	0.32	0.33	***0.96***	0.49
		500	0.73	0.45	0.30	0.29	0.30
		100	0.86	0.62	0.31	0.05	0.08
	Jaccard	1000	0.62	0.33	0.33	***0.96***	0.49
		500	0.73	0.45	0.30	0.29	0.30
		100	0.86	0.62	0.31	0.05	0.08
State-of-the-art link prediction methods	Node2Vec -based	1000	0.74	0.82	0.78	0.81	0.79
		500	0.74	0.70	0.79	0.74	0.76
		100	0.70	0.89	0.78	0.85	0.81
	VGAE	1000	0.80	0.80	0.81	0.78	0.79
		500	0.76	0.78	0.82	0.85	0.83
		100	0.77	0.87	0.83	0.80	0.81
	OLP	1000	0.85	0.52	0.53	0.63	0.58
		500	0.88	0.71	0.59	0.78	0.67
		100	0.89	0.78	0.76	0.82	0.79

Note: The numbers in bold in Dynamic network analysis represent the best values of indicators obtained by our method under different γ values and different scales of dataset. While the numbers in bold in Static network analysis highlight when the data scale was 1000, both PA and Jaccard outperformed our method.

Table 9. Model performance comparison (average value)

Methods	Average indicator values				
	Average AUC	Average accuracy	Average precision	Average recall	Average F1
Our method	0.92	0.88	0.86	0.86	0.85
Common neighbors	0.86	0.80	0.44	0.78	0.53
AA	0.77	0.46	0.28	0.42	0.28
PA	0.74	0.46	0.31	0.43	0.29
Jaccard	0.74	0.47	0.31	0.43	0.29
Node2Vec-based	0.73	0.80	0.78	0.80	0.79
VGAE	0.78	0.82	0.82	0.81	0.81
OLP	0.87	0.67	0.62	0.74	0.68

Expert Evaluation

To further qualitatively evaluate our results, we assembled an expert panel comprising 8 of the top 50 most productive bibliometricians. Each was given the recommendations generated for their own potential collaborators, which included the researcher's name, affiliation, and three of their publications. They were invited to score on each one. The scoring system was on a scale of 0 and 1, where 1 indicates one of the following depending on the circumstance:

– If they have never collaborated, a full willingness to collaborate in the future.
– If they have previously collaborated, they are still willing to maintain the collaboration.

The experts were also given the opportunity to provide additional comments to accompany their scores or to assess the recommendations through written feedback rather than via the scoring system. Five provided scores and three opted for written feedback. Table 10 lists the results of the five experts who scored the recommendations.

Table 10. Results of expert evaluation

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Max	0.8	1	1	1	0.8
Min	0.7	0	0.7	0.1	0.1
Avg	0.73	0.36	0.88	0.52	0.38
Std	0.05	0.45	0.15	0.33	0.26
Mean of the group averages			0.57		

As shown, the maximum score for three of the experts is 1 and 0.8 for the other two, indicating a relatively high satisfaction level with the recommendations. The minimum scores appear to be discrete:

two experts had a minimum score of 0.7, which is still indicative of a strong agreement. Notably, experts' assessments were not only based on the information we provided, but also based on their understandings of the recommended collaborators. Specifically, some experts scored high (greater than or equal to 0.7) because they have similar research interests but had not worked together before. Some of the low scores (less than or equal to 0.5) can be attributed to the fact that these experts considered some practical reasons or did not understand the recommended collaborators.

However, the mean of the scores was only 0.57. While there is room for improvement, an average score of approximately 0.6 is considered to be acceptable for most real-world recommender systems. We also noted that the standard deviation value was rather high in the case of Experts 2 and 5; this result would have been interesting to explore further had they provided comments.

The three experts who provided detailed comments for each recommended collaborator raised some interesting issues about collaboration styles, e.g., interpersonal relationships and working philosophies, which make recommending collaborators a uniquely different proposition from recommending a product to a consumer. One of the experts mentioned that some of the researchers in our recommendations were retired and therefore unlikely to collaborate, and another expert knew the recommended collaborator well; however, there is currently no reason for this expert to collaborate with the recommended author.

Since our experts are leading figures in the field of IS, their assessments are representative. Based on the results of scoring and email comments provided by the experts, we can confirm that our recommendation system has certain practical significance. However, whether scholars cooperate or not often depends on several complex factors. Historical partnerships and similarities in research interests are just a part of these factors. Our recommendation approach cannot consider all the factors associated with the researchers themselves, such as the fact that some may have retired or others may have various interpersonal relationships.

Discussion and Conclusions

This study established a recommendation system for scientific collaborations that can capture deeper insights by considering the changing landscapes in academic collaborations. Incorporated with the ARIMA model, the static network and time-series link prediction method were combined to capture the dynamic evolution of the collaboration between authors. Time-sliced author–keyword matrices were integrated with the Word2Vec model to alleviate the flow of data sparsity in the co-author network, improve the accuracy of similarity measures, and track changes in authors' interests over time. The BSVM, which is a popular ensemble learning model, was incorporated into the framework to improve the performance of potential collaborative recommendations. Quantitative evaluations of the framework on a case study of 22 leading journals in information science between 2005 and 2016 showed that our method has acceptable levels of AUC, accuracy, precision, recall, and F1. A qualitative evaluation was also conducted by a panel of eight experts.

Technical Implications

Link prediction has been widely used in collaboration recommendation systems. However, most studies are limited to simple applications of existing models and indicators that do not reflect the changes in scholarly endeavors and collaboration networks over time. With our framework, the ARIMA model transforms conventional static indicators of future links into dynamic ones, and the experimental results of this study confirm its advantages in collaboration recommendation systems. Such a design can be applied not only to co-author networks, but also to other types of networks to capture topological changes.

The conventional co-word analysis focuses on the collinear relationship of keyword frequency. In our framework, the Word2Vec model is applied to reduce the noise between different terms but with similar meanings, and the semantic and contextual relationships between keywords are fully considered. This makes the similarity measures of authors' research interests more accurate. Additionally, the Word2Vec model is suitable for large-scale datasets. Therefore, we believe that our method will be effective for large-scale semantic analysis.

Possible Applications

Three applications of this framework are immediately apparent:

1) Our method can not only provide services for individual researchers to find academic partners, but also help scientific funding agencies and research institutions to conduct top-level layouts and organize and guide researchers to achieve strong alliance.
2) By simply replacing the corpus of academic papers with other ST&I data, such as patents or articles in industry journals, the framework could become a recommendation engine for enterprises seeking cooperation partners or, conversely, as a competitor analysis tool.
3) As a method integrating topological analysis with semantic analysis, our framework is suitable for performing many other bibliometric tasks via network-based analysis, such as identifying emerging technologies or suggesting research topics to academics who wish to change or extend their fields of interest.

Limitations and Future Studies

Some of the limitations of the current research and related future directions are summarized as follows: 1) Selecting collaborators is a complex decision-making process that is influenced by many factors: geography, organization culture, working style, and gender. Developing a quantitative system that considers all these factors is both a challenge and opportunity—and one that we leave to time and future endeavors of researchers in this field, including ourselves; 2) Currently, link prediction methods can only predict connections that might appear in the future; they cannot predict those that might disappear. An interesting direction for future research may be to consider profiles of academic activity as an indicator of when collaborations or academic careers might be nearing their end; 3) Our focus in the case study was limited to information science and, therefore, only articles published in information science-related journals were included in the corpus. Nevertheless, it is reasonable to assume that at least some of these authors have interests in other fields and have therefore published papers in journals that were not included in our dataset. A promising next step may be to use a global dataset to test our method and explore empirical insights for multiple disciplines and research areas; 4) This paper emphasizes the accuracy with which scientific collaborators can be recommended, rather than the efficiency of the recommendation system. In addition, keyword processing and network construction are complex and may be time-consuming. In the future, we hope to leverage multithreading to improve the efficiency of the current recommendation system.

Acknowledgments

This work was supported by the National Science Foundation of China Funds [Grant No. 71774013] and the Australian Research Council under Discovery Early Career Researcher Award DE190100994. Our heartfelt appreciation goes to Xiaoli Cao and Changtian Wang for their contributions to this paper.

References

Adamic, L. A. & Adar, E. (2003). Friends and neighbors on the Web. Social Networks, 25, 211–230.

Akaike, H. T. (1974). A New Look at the Statistical Model Identification. Automatic Control IEEE Transactions on Automatic Control, 19(6), 716-723.

Alinani, K., Alinani, A., Narejo, D. H., & Wang, G. (2018). Aggregating author profiles from multiple publisher networks to build a list of potential collaborators. IEEE Access, 6, 20298-20308.

Benchettara, N., Kanawati, R., & Rouveirol, C. (2010). A supervised machine learning link prediction approach for academic collaboration recommendation. Proceedings of the fourth ACM conference on Recommender systems (pp. 253-256). Berlin, Heidelberg: Springer.

Bian, J., Xie, M., Topaloglu, U., Hudson, T., Eswaran, H., & Hogan, W. (2014). Social network analysis of biomedical research collaboration networks in a CTSA institution. Journal of biomedical informatics, 52, 130-140.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Box, G. E., & Jenkins, G. M. (1976). Time series analysis: forecasting and control rev. ed. Oakland, California, Holden-Day, 31(4), 238-242.

Chaiwanarom, P., & Lursinsap, C. (2015). Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status. Knowledge-Based Systems, 75, 161–172.

Chen, R., Hua, Q., Chang, Y. S., Wang, B., Zhang, L., & Kong, X. (2018). A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. IEEE Access, 6, 64301-64320.

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).

Chen, W., Qu, H., & Chi, K. (2021). Partner Selection in China Interorganizational Patent Cooperation Network Based on Link Prediction Approaches. Sustainability, 13(2), 1003.

Cho, H., & Yu, Y. (2018). Link prediction for interdisciplinary collaboration via co-authorship network. Social Network Analysis and Mining, 8(1), 1-12.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (2010). Indexing by latent semantic analysis. Journal of the Association for Information Science & Technology, 41(6), 391-407.

Eck, N. J. V., & Waltman, L. (2008). Appropriate Similarity Measures for Author Cocitation Analysis. Journal of the Association for Information Science & Technology, 59(10), 1653-1661.

Eck, N. J. V., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 523-538.

Fawcett, T. (2005). An introduction to roc analysis. Pattern Recognition Letters, 27(8), 861-874.

Gao, Z., Fan, Y., Wu, C., Tan, W., Zhang, J., Ni, Y., ... & Chen, S. (2018). Seco-lda: Mining service co-occurrence topics for composition recommendation. IEEE Transactions on Services Computing, 12(3), 446-459.

Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airoldi, E & Clauset, A. (2019). Stacking Models for Nearly Optimal Link Prediction in Complex Networks.

Grover, A. & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. KDD: proceedings. International Conference on Knowledge Discovery & Data Mining. 2016. 855-864.

Gui, Q., Liu, C., & Du, D. B. (2019). The structure and dynamic of scientific collaboration network

among countries along the belt and road. Sustainability, 11(19), 5187.
Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. Scientometrics, 101(2), 1461-1473.
Guo, G., Zhang, J., & Yorke-Smith, N. (2016). A novel recommendation model regularized with user trust and item ratings. ieee transactions on knowledge and data engineering, 28(7), 1607-1620.
Haruna, K., Ismail, M. A., Damiasih, D., Sutopo, J., & Herawan, T. (2017). A collaborative approach for research paper recommender system. PLoS ONE, 12(10), e0184516.
He, C., Wu, J., & Zhang, Q. (2021). Proximity‐aware research leadership recommendation in research collaboration via deep neural networks. Journal of the Association for Information Science and Technology.
Hoang, D. T., Nguyen, N. T., Tran, V. C., & Hwang, D. (2019). Research collaboration model in academic social networks. Enterprise Information Systems, 13(7-8), 1023-1045.
Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: a document co-citation analysis (2009–2016). Scientometrics, 115(2), 869-892.
Hu, K., Wu, H., Qi, K., Yu, J., Yang, S., Yu, T., Zheng, J., & Liu, B. (2018). A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. Scientometrics, 114(3), 1031-1068
Huang, L., Zhu, Y., Zhang, Y., Zhou, X., & Jia, X. (2018). A link prediction-based method for identifying potential cooperation partners: A case study on four journals of informetrics. In 2018 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 1-6). IEEE.
Huang, Z., & Lin, D. K. J. (2009). The time-series link prediction problem with applications in communication surveillance. Informs Journal on Computing, 21(2), 286-303.
Işık, G. T. Z. (2018). A Hybrid Movie Recommendation System Using Graph-Based Approach. International Journal of Computing, 7(2), 29-37.
Jeh, G., & Widom, J. (2002). SimRank: A Measure of Structural-Context Similarity. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 538-543). New York: ACM Press.
Junuthula, R. R., Xu, K. S., & Devabhaktuni, V. K. (2016). Evaluating link prediction accuracy in dynamic networks with added and removed edges. In 2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom) (BDCloud-SocialCom-SustainCom) (pp. 377-384). IEEE.
Kim, H. C., Pang, S., Je, H. M., Kim, D., & Bang, S. Y. (2002a). Pattern Classification Using Support Vector Machine Ensemble (pp. 160-163). International Conference on Pattern Recognition. IEEE.
Kim, H. C., Pang, S., Je, H. M., Kim, D., & Bang, S. Y. (2002b). Support vector machine ensemble with bagging. Pattern Recognition with Support Vector Machines, 2388, 397-407.
Kim, J., & Diesner, J. (2019). Formational bounds of link prediction in collaboration networks. Scientometrics, 119(2), 687-706.
Kipf, T. & Welling, M. (2016). Variational Graph Auto-Encoders.
Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., & Tolba, A. (2016). Exploiting publication contents and collaboration networks for collaborator recommendation. PloS One, 11(2), e0148492.
Kong, X., Mao, M., Wang, W., Liu, J., & Xu, B. (2018). VOPRec: Vector representation learning of papers with text information and structural identity for recommendation. IEEE Transactions on

Emerging Topics in Computing.
Lai, T. L., & Tsang, K. W. (2016). MULTIVARIATE STOCHASTIC REGRESSION IN TIME SERIES MODELING. Statistica Sinica, 1411-1426.
Lande, D., Fu, M., Guo, W., Balagura, I., Gorbov, I., & Yang, H. (2020). Link prediction of scientific collaboration networks based on information retrieval. World Wide Web, 1-19.
Lee, D. H., Brusilovsky, P., & Schleyer, T. (2011). Recommending collaborators using social features and MeSH terms. Proceedings of the American Society for Information Science and Technology, 48(1), 1-10.
Lee, D.H., & Schleyer, T. (2010). MeSH term explosion and author rank improve expert recommendations. AMIA Annu Symp Proc, 13, 412-6.
Leydesdorff, L. (2005). Similarity measures, author cocitation analysis, and information theory. Journal of the American Society for Information Science and Technology, 56(7), 769-772.
Li, J. C., Zhao, D. L., Ge, B. F., Yang, K. W., & Chen, Y. W. (2018a). A link prediction method for heterogeneous networks based on BP neural network. Physica A: Statistical Mechanics and its Applications, 495, 1-17.
Li, Z., Zhang, H., Wang, S., Huang, F., Li, Z., & Zhou, J. (2018b). Exploit latent Dirichlet allocation for collaborative filtering. Frontiers of Computer Science, 12(3), 571-581.
Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. Journal of the Association for Information Science and Technology, 58(7), 1019-1031.
Liu, Y., Wang, S., Khan, M. S., & He, J. (2018). A novel deep hybrid recommender system based on auto-encoder with neural collaborative filtering. Big Data Mining and Analytics, 1(3), 211-221.
Lü, L., & Zhou, T. (2011). Link prediction in complex networks: a survey. Physical A Statistical Mechanics & Its Applications, 390(6), 1150-1170.
Lu, Z. G., & Chen, Q. (2020). Discovering Potential Partners via Projection-Based Link Prediction in the Supply Chain Network. International Journal of Computational Intelligence Systems, 13(1), 1253-1264.
Mcinnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of Biomedical Informatics, 46(6), 1116-1124.
Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), 1-12.
Meyffret, S., Médini, L., & Laforest, F. (2013). Confidence on collaborative filtering and trust-based recommendations. In International Conference on Electronic Commerce and Web Technologies (pp. 162-173). Springer, Berlin, Heidelberg.
Mikolov, T., Chen, K., Corrado, G., &Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Computer Science. arXiv preprint arXiv:1301.3781, 2013.
Molaei, S., Havvaei, A., Zare, H., & Jalili, M. (2021). Collaborative Deep Forest Learning for Recommender Systems. IEEE Access, 9, 22053-22061.
Newman, M.E.J. (2001). Clustering and preferential attachment in growing networks. Working Papers, 64(2), 025102.
Pradhan, T., & Pal, S. (2020). A multi-level fusion based decision support system for academic collaborator recommendation. Knowledge-Based Systems, 197, 105784.
Schreiber, M., Kraft, B., & Zündorf, A. (2017). Metrics Driven Research Collaboration: Focusing on Common Project Goals Continuously (pp. 41-47). IEEE/ACM 4th International Workshop on Software Engineering Research and Industrial Practice (SER and IP).

Shani, G., Meek, C., & Gunawardana, A. (2001). Hierarchical Probabilistic Segmentation of Discrete Events. Intelligent Data Analysis,15(4),483-501.

Shibata, N., Kajikawa, Y., & Sakata, I. (2011). Link prediction in citation networks. Journal of the Association for Information Science & Technology, 63(1), 78-85.

Sugiyama, K., Kumar, T., Kan, M. Y., & Tripathi, R. C. (2010). Identifying citing sentences in research papers using supervised learning (pp. 67-72). 2010 International Conference on Information Retrieval & Knowledge Management (CAMP). IEEE.

Sun, C., Ng, K. T. J., Henville, P., & Marchant, R. (2018). Hierarchical Word Mover Distance for Collaboration Recommender System. In Australasian Conference on Data Mining (pp. 289-302). Springer, Singapore.

Sun, N., Lu, Y., & Cao, Y. (2019). Career age-aware scientific collaborator recommendation in scholarly big data. IEEE Access, 7, 136036-136045.

Terveen, L., & McDonald, D. W. (2005). Social matching: A framework and research agenda. ACM transactions on computer-human interaction (TOCHI), 12(3), 401-434.

Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for computer science publications. Knowledge-Based Systems, 157, 1-9.

Wang, M., Wu, Z., Sun, X., Feng, G., & Zhang, B. (2019a). Trust-aware collaborative filtering with a denoising autoencoder. Neural Processing Letters, 49(2), 835-849.

Wang, W., Feng, Y., Jiao, P., & Yu, W. (2017). Kernel framework based on non-negative matrix factorization for networks reconstruction and link prediction. Knowledge Based Systems, 137(1), 104-114.

Wang, W., Xu, B., Liu, J., Cui, Z., Yu, S., Kong, X., & Xia, F. (2019b). CSTeller: forecasting scientific collaboration sustainability based on extreme gradient boosting. World Wide Web, 22(6), 2749-2770.

Wang, Y., Liu, Z., & Sun, M. (2015). Incorporating linguistic knowledge for learning distributed word representations. PLOS ONE, 10.

White, H., & Mccain, K. (2010). Visualizing a discipline: an author co-citation analysis of information science, 1972–1995. Journal of the Association for Information Science and Technology, 49(4), 327–355.

Xia, H., Luo, Y., & Liu, Y. (2021). Attention neural collaboration filtering based on GRU for recommender systems. Complex & Intelligent Systems, 1-13.

Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. Journal of Informetrics, 8(2), 295-309.

Yang, C., Liu, T., Chen, X., Bian, Y., & Liu, Y. (2020). HNRWalker: recommending academic collaborators with dynamic transition probabilities in heterogeneous networks. Scientometrics, 1-21.

Zhang, C., Wu, X., Yan, W., Wang, L., & Zhang, L. (2019a). Attribute-aware graph recurrent networks for scholarly friend recommendation based on internet of scholars in scholarly big data. IEEE Transactions on Industrial Informatics, 16(4), 2707-2715.

Zhang, H., Kong, X., & Zhang, Y. (2021). Selective Knowledge Transfer for Cross-Domain Collaborative Recommendation. IEEE Access, 9, 48039-48051.

Zhang, Q., Mao, R., & Li, R. (2019b). Spatial–temporal restricted supervised learning for collaboration recommendation. Scientometrics. 119:1497-1517.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. Journal of Informetrics, 12(4), 1099-1117.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). “Term clumping” for technical intelligence: A case study on dye-sensitized solar cells. Technological Forecasting and Social Change, 85, 26-39.

Zhang, Y., Zhang, C., & Liu, X. (2017a). Dynamic Scholarly Collaborator Recommendation via Competitive Multi-Agent Reinforcement Learning. the Eleventh ACM Conference. ACM.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017b). Scientific evolutionary pathways: identifying and visualizing relationships for scientific topics. Journal of the Association for Information Science and Technology, 68(8), 1925-1939.

Zhao, D., & Strotmann, A. (2010). Evolution of research activities and intellectual influences in information science 1996-2005: introducing author bibliographic-coupling analysis. Journal of the American Society for Information Science and Technology, 59(13), 2070-2086.

Zhou, P., & Jiang, D. (2020). Study on the evolution of hot topics in the urban development. Evolutionary Intelligence, 1-9.

Zhou, T., Lü, L., & Zhang, Y. C. (2009). Predicting missing links via local information. The European Physical Journal B, 71(4), 623-630.

Zhou, X., Ding, L., Li, Z., & Wan, R. (2017). Collaborator recommendation in heterogeneous bibliographic networks using random walks. Information Retrieval Journal, 20(4), 317-337.

Zhu, T. T., & Lan, M. (2014). Measuring short Text Semantic Similarity using multiple measurements. International Conference on Machine Learning and Cybernetics, 1(4), 808-813.