World Wide Web
https://doi.org/10.1007/s11280-019-00768-9

Link prediction of scientific collaboration networks based on information retrieval

Dmytro Lande, et al. *[full author details at the end of the article]*

Received: 24 December 2018 / Revised: 20 November 2019 / Accepted: 5 December 2019
Published online: 07 March 2020

Abstract
Link prediction plays an important role in scientific collaboration networks, and can favourably affect the organization of international scientific projects. In this paper, a meta-path computed prediction (MPCP) algorithm for link prediction among scientists and publications is presented. The MPCP algorithm is based on a heterogeneous information network model composed of authors and keywords in articles retrieved from the Web of Science database. Two kinds of meta-paths are defined: Author to Author to Author (A-A-A) and Author to Direction to Author (A-D-A). By calculating A-A-A and A-D-A using the heterogeneous information network model, the predictive strength of the links can be computed. The overlap of the meta-paths is also taken into account. By restoring links and calculating the number of restored links with different standard values, similar results are achieved for (quantum communication and link prediction). The number of restored links decreases as a special threshold value increases. The experimental studies show that, for any threshold value up to 1, at least 50% of links are restored. The results presented in this paper verify that the algorithm is a feasible means of predicting collaboration among scientists.

Keywords link prediction · scientific collaboration network · meta-path · random walk

1 Introduction

The most ambitious scientific projects, such as the relativistic heavy ion collider and the international space station, usually involve cooperation among a large number of scientists. Nowadays, scientists from different countries join together in developing projects and research that contribute to worldwide progress. Thus, it is important to organize common scientific

Highlights
- An algorithm for link prediction among scientists and publications is proposed.
- Importance of using heterogeneous information networks is discussed.
- Advantages of the proposed algorithm are confirmed via examples.
- Perspectives of using the proposed algorithm are discussed.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11280-019-00768-9) contains supplementary material, which is available to authorized users.

research via scientific collaboration. Scientific collaboration networks are used to describe the relationship between scientific collaborators on the basis of co-authorship [1]. One of the earliest studies of the scientific collaboration network was conducted by the famous mathematician Pául Erdös. The cooperation between Erdös and 500 scientists formed a huge scientific collaboration network [2].

Scientific collaboration networks have the small-world phenomenon, which obey a power-law distribution [3, 4]. In the early days of scientific collaboration network research, the main topics concerned the nature of scientific collaboration networks. Scientists investigated centrality indices, clustering, behaviour distribution, and other features of scientific collaboration networks [3]. The evolution of collaboration networks was also widely studied [5, 6]. Abbasi et al. studied the features of growing networks with respect to centrality measures, and described the correlation between authors' centrality measures and the attachment frequency of new authors to them. They concluded that the betweenness centrality could be used to predict the preferential attachment of new nodes [6]. Milojević proposed a model that explained the principles underlying the formation and evolution of scientific research teams [7]. These findings were common features of scientific collaboration networks.

Recently, link prediction has been an important issue in the study of scientific collaboration networks. Kenekayoro et al. investigated whether using machine learning methods to filter page types could improve the extent to which hyperlink data can be used to indicate the extent of collaboration between universities [8]. Link prediction aimed to predict missing links in current networks and new or dissolution links in future networks, which are important for mining and analysing the evolution of social networks [9]. To improve the predictive performance, some scientists proposed some new methods [10–13]. For example, Ghasemian et al. used the information in collaborative networks to extract features that improve the predictive performance. Sett et al used a robust and efficient feature set called TMLP (Time-aware Multi-relational Link Prediction) for link prediction in dynamic heterogeneous networks [14]. Symeonidis et al defined a basic node similarity measure and exploited global graph features introducing transitive node similarity to optimize the recommendation algorithm [15]. For another, the study of link structure played an important role in the network [16, 17].

Kenekayoro et al. investigated whether using machine learning methods to filter page types could improve the extent to which hyperlink data can be used to indicate the extent of collaboration between universities [8]. Gleich reviewed Google's PageRank method, which was developed to evaluate the importance of Webpages via their link structure [9], and He et al. proposed a link prediction ensemble algorithm to obtain more stable prediction performance [10]. Link prediction aimed to predict missing links in current networks and new or dissolution links in future networks, which are important for mining and analysing the evolution of social networks [11]. Lue et al. proposed a universal structural consistency index based on the perturbation of the adjacency matrix that requires no prior knowledge of the network organization [12], while Ghasemian et al. used the information in collaborative networks to extract features that improve the predictive performance [13]. Li et al. used an Expectation–Maximization algorithm to estimate certain parameters and form predictions based on utility analysis [14], whereas Valverde-Rebaza et al. proposed a friendship prediction method based on location [15]. Kefalas et al. provided a novel recommendation method based on the time dimension [16]. Sett et al used a robust and efficient feature set called TMLP (Time-aware Multi-relational Link Prediction) for link prediction in dynamic heterogeneous networks [17]. Symeonidis et al defined a basic node similarity measure and exploited global graph features introducing transitive node similarity to optimize the recommendation algorithm [18]. Authors of paper [19] proposed two novel node-coupling clustering approaches and their extensions for link prediction, which combine the coupling degrees of the common neighbour nodes of a predicted node-pair with cluster geometries of nodes. Muhan Zhang and Yixin Chen develop a novel -decaying heuristic theory for link prediction based on graph neural networks [20]. Haeran Cho and Yi Yu proposed a new link prediction methodology, with the specific aim of identifying potential interdisciplinary collaboration in a university-wide collaboration network [21].

Link prediction is the determination of some linkage between two nodes in a network that have not yet been connected by known nodes and structures. Some of the above studies are based on nodes, whereas others are based on the network structure. We think that scientific collaboration networks contain many types of nodes and edges. So, we consider scientific collaboration networks to be heterogeneous information networks in this paper. Hence, we propose a meta-path computed prediction (MPCP) algorithm based on meta-paths and random walks that predicts cooperative relationships. We combine the meta-paths with random walks, and take the overlaps of the meta-paths into account. Through the restoration of links and determination of the number of restored links with different standard values, we demonstrate the excellent performance of the proposed algorithm.

2 Heterogeneous information network model for scientific collaboration network

2.1 Scientific collaboration network based on co-authors and keywords

In this paper, we address the issue of scientific collaboration networks from the viewpoint of literature retrieval. Nowadays, most research groups across the world would like to share their findings by means of publishing papers. Thus, to some extent, the literature in a particular field, especially published papers, can reveal the intuitive relationships within scientific collaboration networks.

We used Clarivate Analytics' Web of Science as the literature retrieval database [22] and selected quantum communication as the topic. We retrieved 99 literature items published from January 1st–December 31st, 2017, and constructed a co-authors network and keywords network using the VOSviewer software [23], as shown in Figs. 1 and 2.

Figure. 1 shows a network of co-authors that published articles related to quantum communication in 2017. The nodes represent authors and the edges represent common papers by these authors. The bigger the node, the more articles the author has published concerning quantum communication. The scientific collaboration network assigns nodes to several clusters describing scientific collaboration communities [24]. In the figures, clusters are indicated by different colours. For example, in the biggest community, Prof. J.W. Pan made a significant contribution to quantum communications in 2017, and the network shows his widespread cooperation with other communities.

The relations among the keywords of the 99 papers are shown in Fig. 2, where each dot represents a keyword. The size of the dot corresponds to the frequency with which this keyword was used. According to the frequency of simultaneous use, the keywords are displayed in different colours. The same colour represents the higher frequency of simultaneous use. The two keywords most associated with quantum communication in 2017 were 'quantum key distribution' and 'entanglement concentration'.

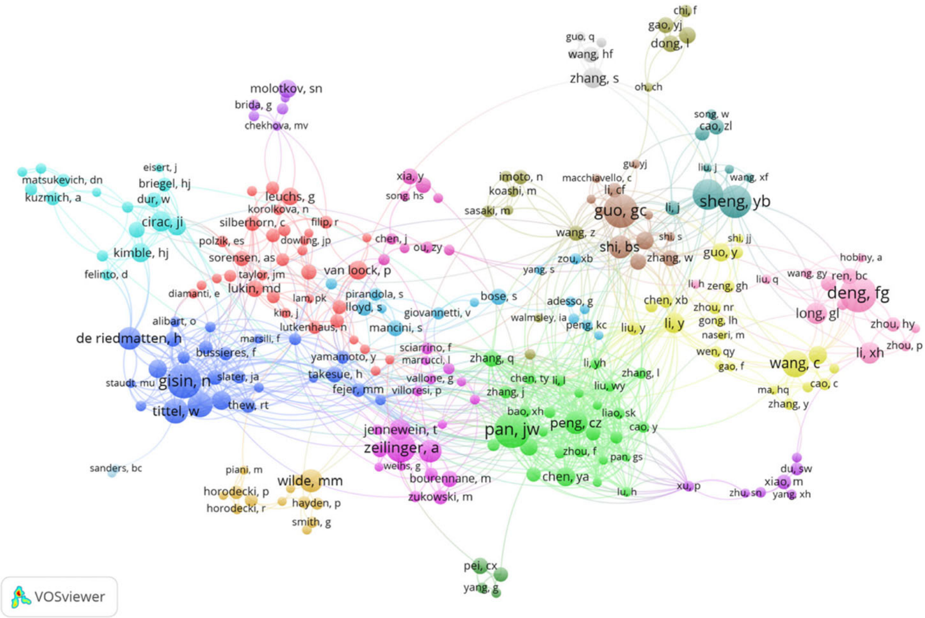

Fig. 1 Co-authorship network of quantum communication built by VOSviewer. Different colours represent different research groups

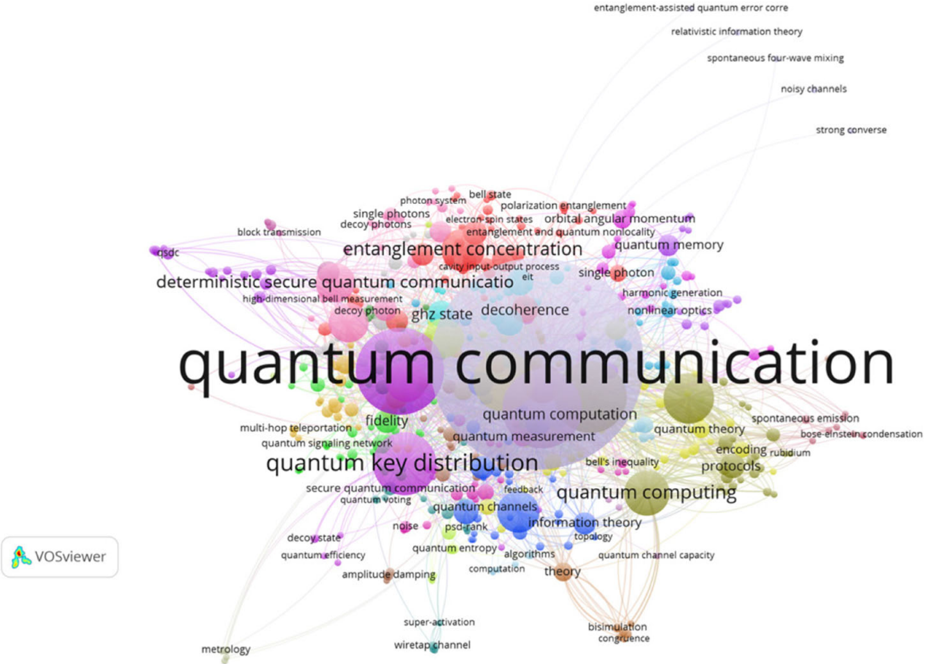

Fig. 2 Keywords network of quantum communication built by VOSviewer. Different colours represent different keywords

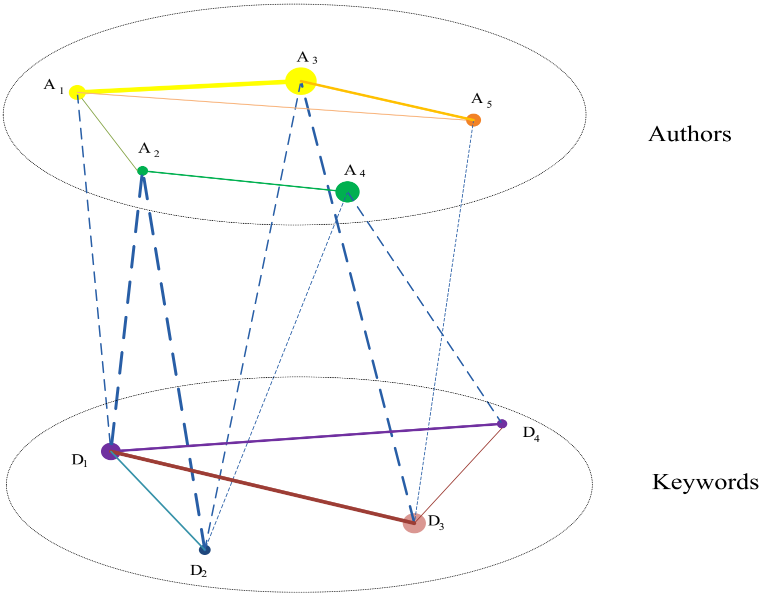

Fig. 3 Heterogeneous network structure model for authors and keywords. The upper layer represents the co-authorship between authors and the bottom layer represents the relationship between different keywords. The link between the two layers represents the relationship between authors and keywords

Figures. 1 and 2 demonstrate how a scientific collaboration network can be established by means of a literature retrieval on the subject of quantum communication. We can acquire some useful information about research groups and research hotspots from Figs. 1 and 2, respectively. However, the scientific collaboration networks in Figs. 1 and 2 can be classified as homogeneous information networks [25, 26]—there is a strong dependence between them. Actually, if we compose the scientific collaboration networks in Figs. 1 and 2 as heterogeneous information networks, we could acquire more useful information [27–29]. Moreover, it is more reasonable to make link predictions according to the heterogeneous information networks. For example, Ma et al. built dynamic heterogeneous information networks to predict neighbour label distributions [29], while Ozcan et al. proposed a novel multivariate method for link prediction in evolving heterogeneous networks using a nonlinear autoregressive neural network with external inputs [30]. Li et al. developed a novel integrated framework called the meta-path feature-based backpropagation neural network model to predict multiple link types for heterogeneous networks [31].

2.2 Heterogeneous information network model

In this paper, we define a heterogeneous network structure model as $G = (A, D, R)$. This is a directed network in which $A = \{A_1, A_2, \ldots, A_n\}$ is a set of nodes representing authors and $D = \{D_1, D_2, \ldots, D_n\}$ is a set of nodes representing research directions. We assume that the research directions can be described by the keywords extracted from the papers and can be treated as the authors' research interests. Assume that $R = \{R_{AA}, R_{DD}, R_{AD}, R_{DA}\}$ is a set of edges, where $R_{AA} = \{(A_i, A_j) | A_i, A_j \in A\}$ denotes the co-authorship between authors A_i and A_j, $R_{DD} = \{(D_i, D_j) | D_i, D_j \in D\}$ describes connections between keywords D_i and D_j, and $R_{DA} = \{(D_i, A_j) | D_i \in D,\ A_j \in A\}$ and $R_{AD} = \{(A_i, D_j) | A_i \in A,\ D_j \in D\}$ are the relationships between authors and keywords.

2.3 MPCP algorithm

2.3.1 Meta-path

Let meta-path T_i set $T_i = S_1 \xleftrightarrow{R_1} S_2 \xleftrightarrow{R_2} \cdots \xleftrightarrow{R_{n-1}} S_n$, where S_i∈A∪D and R_j(j = 1, ..., $n - 1$)is the links between nodes of different types [32, 33]. The main idea of link prediction is based on meta-paths. Paths containing three or fewer hops (or degrees) are considered to be strong [34]; otherwise, the path is weak. Strong ties indicate frequent working partners, whereas weak ties suggest partners with fewer opportunities for cooperation. We discuss two meta-paths denoting strong ties.

Case 1: A-A-A. Authors may establish partnerships with collaborators or co-authors. In Fig. 4 links between authors of A-A-A type are presented. We made an assumption about connection of type A_1– A_2 – A_3, using the fact that the authors A_1 and A_2 are connected, as well as A_2 and A_3.

Case 2: A-D-A. Authors may establish relationships with people who publish papers with similar keywords. The links of type A-D-A between authors and common key-word are shown on the fig. 5. We assumed the links of type A_1– D_1 – A_2 because author A_1 published a paper with the keyword D_1, and A_2 also published a work with the same keyword.

2.3.2 MPCP model

The nodes in the network are assumed to be independent, as are the links. We predict the probability of cooperation between authors based on a random walk [35, 36].

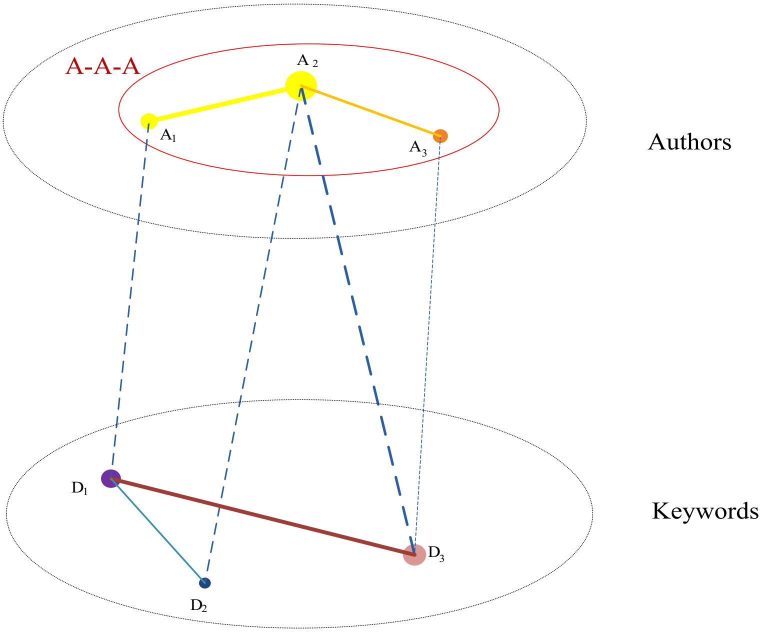

Fig. 4 Meta-path: A-A-A

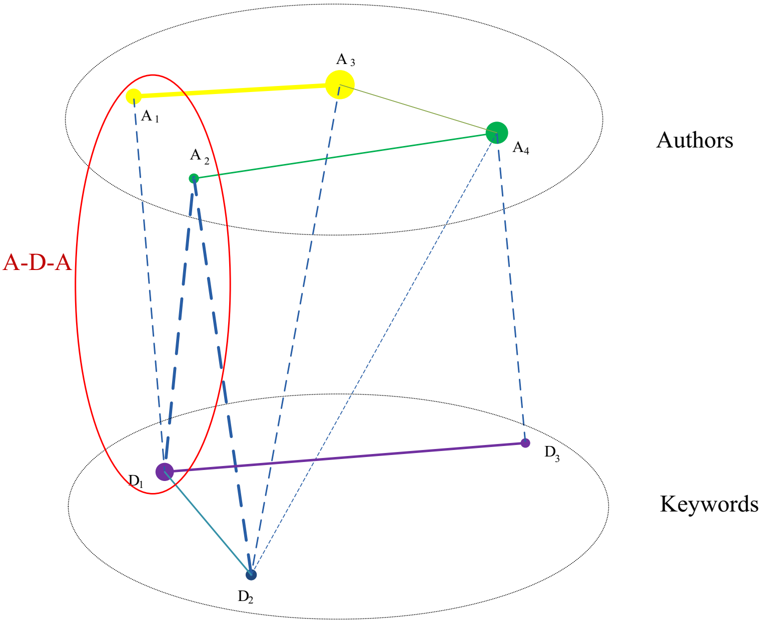

Fig. 5 Meta-path: A-D-A

Let P represent the node connection probability matrix in the structural network:

$$\mathrm{P} = \begin{bmatrix} P_{AA} & P_{AD} \\ P_{DA} & P_{DD} \end{bmatrix} \tag{1}$$

where P_{AA} is an $N \times N$ matrix representing the relationships between authors. If there is no cooperation, the elements will be equal to 0.

Element P_{a1a2} of matrix P_{AA} is calculated as:

$$p_{a_1 a_2} = \frac{N\left(a_1 \cap a_2\right)}{N(a_1) + N(a_2) - N\left(a_1 \cap a_2\right)}, \ \mathit{if} \ N(a_1)N(a_2) > 0, \ \mathit{else} \ p_{a_1 a_2} = 0 \tag{2}$$

In this expression, *N(a*$_1$*)* is the number of first-author's papers, *N(a*$_2$*)* is the number of second-author's papers, and *N(a*$_1$*∩a*$_2$*)* is the number of common papers written by both authors.

The elements P_{ad} of the $N \times M$ matrix P_{AD} are given by:

$$p_{ad} = \frac{N(a, d)}{N(a)}, \ \mathit{if} \ N(a) > 0, \ \mathit{else} \ p_{ad} = 0 \tag{3}$$

Here, *N(a,d)* is the number of papers published by author *a* in direction *d*, and *N(a)* is the number of papers published by author *a*.

The elements P_{da} of the $M \times N$ matrix P_{DA} are calculated as:

$$p_{da} = \frac{N(a, d)}{N(d)}, \ \mathit{if} \ N(d) > 0, \ \mathit{else} \ p_{da} = 0. \tag{4}$$

In this equation, *N(a,d)* is the number of papers published by author *a* in direction *d* and *N(d)* is the number of papers published in direction *d*.

The elements P_{d1d2} of matrix P_{DD} are calculated by:

$$p_{d_1 d_2} = \frac{N\left(d_1 \cap d_2\right)}{N(d_1) + N(d_2) - N\left(d_1 \cap d_2\right)}, \ \textit{if} \ N(d_1)N(d_2) > 0, \ \textit{else} \ p_{d_1 d_2} = 0. \tag{5}$$

In this case, $N(d_1)$ is the number of papers published in direction d_1, $N(d_2)$ is the number of papers published in direction d_2, and $N(d_1 \cap d_2)$ is the number of papers in both directions.

Most papers have several keywords. We deem keywords to be similar if they appear in the same paper. Thus, $p_{d_1,d_2} = 1$ when d_1 and d_2 appear in the same paper, and $p_{d_1,d_2} = 0$ otherwise.

The proposed method calculates the predictive force of links among pairs of authors by taking into account their possible links through other authors (Fig. 4, A-A-A) or through keywords (Fig. 5, A-D-A). The predictive strength of a link is given by:

$$f_{a_i a_j} = 1 - \left(1 - p_{a_i a_j}\right) H_{ij}, \tag{6}$$

where $f_{a_i a_j}$ denotes the force of links among pairs of authors a_i and a_j, and $p_{a_i a_j}$ is the probability of a direct connection between authors and H_{ij} defined as:

$$H_{ij} = \begin{cases} \prod_{k=1}^{K} \left(1 - p_{a_i t_k} p_{t_k a_j}\right) & K > 0 \\ 1 & K = 0 \end{cases}. \tag{7}$$

where $t_k \in A \cup D$ are intermediate nodes (authors or keywords). Note that $t_k \neq a_i$ and $t_k \neq a_j$. K is the number of intermediate nodes between a_i and a_j.

In the paper, we set an empirical threshold $\tau \in [0, 1]$ for the stepwise removal of links with predictive strength of less than 0.1. Set an empirical α - threshold of restoring links, set $\alpha = 0.5$ in all experiments.

To assess the possibility of restoring connections in the network after they have been destroyed, the following procedure is proposed:

1. Build an initial network for a sample set of documents. Relations are calculated and normalized to 1 according to the formulas presented above, following matrix (1).
2. For this purpose, links with a weight not less than a threshold τ are removed from the constructed network.
3. Recalculate new edge weights using formula (6).
4. Calculate the proportion of restored links.

2.4 MPCP algorithm

The proposed prediction algorithm consists of six main steps, which can be described as follows:

Step 1: Collect author/keyword data from Web of Science.

Step 2: Form a heterogeneous information network G = (A,D,R). Authors and keywords are used to establish the two layers. The connection weights between the two layers are determined according to the number of cooperative relationships.

Step 3: Compute A-A-A and A-D-A meta-paths according to G and generate the meta-path set T.

Step 4: Links with a weight not less than a threshold (τ) are removed from the constructed network.

Step 5: Using the proposed algorithm, calculate the connectional strength using Eq. (6). If the strength of the predictive connection is greater than α =0.5, the link is restored.

Step 6: Calculation of parameters:

Completeness - proportion of restored links equal to the ratio of the number of correctly restored links to the total number of deleted links; accuracy equal to the ratio of the number of correctly restored links to the total number of restored links.

Step 7: After restoring of all necessary links, the network of authors will be drawn. Output of the result about the percentage of restored links.

The time complexity of the presented algorithm the time of the most complex step 3 and is limited by the values $O(N_A^3)$ with $N_A \geq N_D$ and $O(N_A^3\ N_D)$ with $N_A < N_D$. Where N_A denotes the number of authors (size of the set A), and N_D denotes the size of the set D.

The pseudocode of the MPCP algorithm is as follows:

```
Begin.
READ the source dataset.
FORM the matrix (1).
For $\tau$ = 0 to 1 do.
Remove connections with weights not bigger than $\tau$.
Recalculate the weights of edges using formula (6).
Determine parameters: the proportion of restored links and the Build a graph, display the values.
End For.
End
```

3 Simulation results and discussion

3.1 Data collection

We obtained two datasets from the Web of Science [22]. The keywords used to extract the datasets were quantum communication and link prediction, respectively. We collected 99 articles on quantum communication and 132 articles about link prediction published in 2017.

3.2 Simulation results

On a Windows 10 machine, we used a multi-platform Perl language for the subsequent visualization of Gephi [37]. To construct the author connections subgraph, the distribution graph of the node degrees was built. The obtained distribution was compared with the work of Newman [38], and was found to obey a power law for which the recovery percentages are known.

If the strength of the predictive connection between authors (calculated by Eq. (6)) was sufficiently large, i.e. greater than 0.5, there was assumed to be a relationship between the authors, even if this was not determined in advance by direct data collection. In both the original networks, the most powerful connection edges (having a weight of 1) were removed. Using the proposed algorithm, we calculated the predictive relationships and the percentage of bonds thus restored. Note that connections that have not been determined in advance can also be ‘restored’. This is not an error of the algorithm. In this way, connections that were not originally taken into account when collecting the data can be re-established.

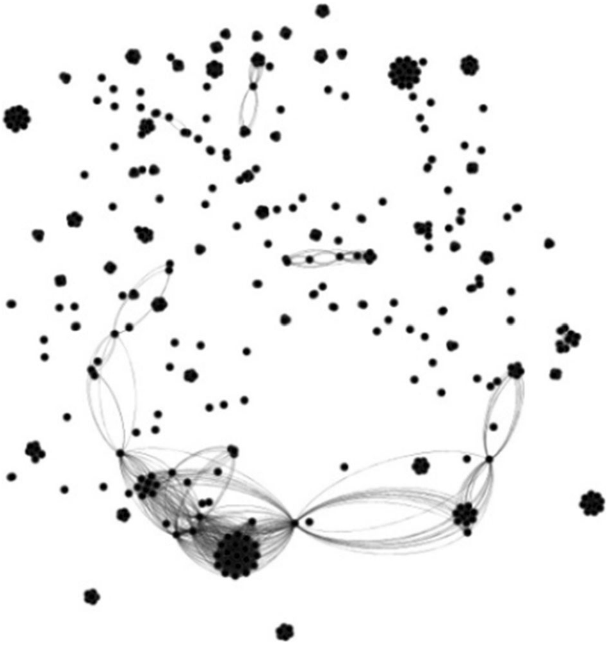

Fig. 6 Whole network – source graph with 2314 edges, density 0.02, average degree 6.7, and 63 connected components

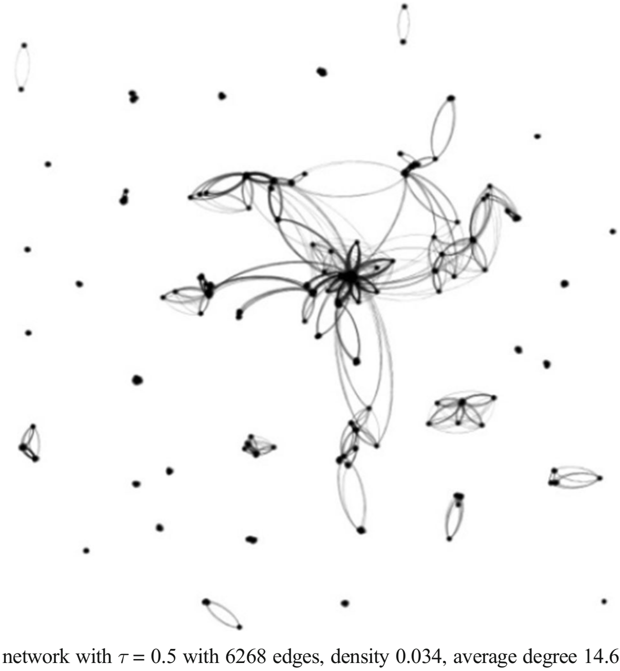

Fig. 7 Predictive network with $\tau = 0.5$ with 6268 edges, density 0.034, average degree 14.6, and 31 connected components

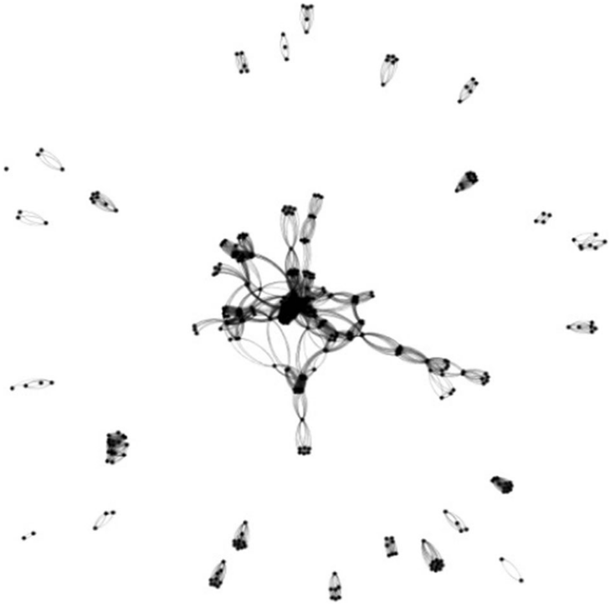

Fig. 8 Predictive network with $\tau = 0.4$ with 7495 edges, density 0.04, average degree 14.65, and 29 connected components

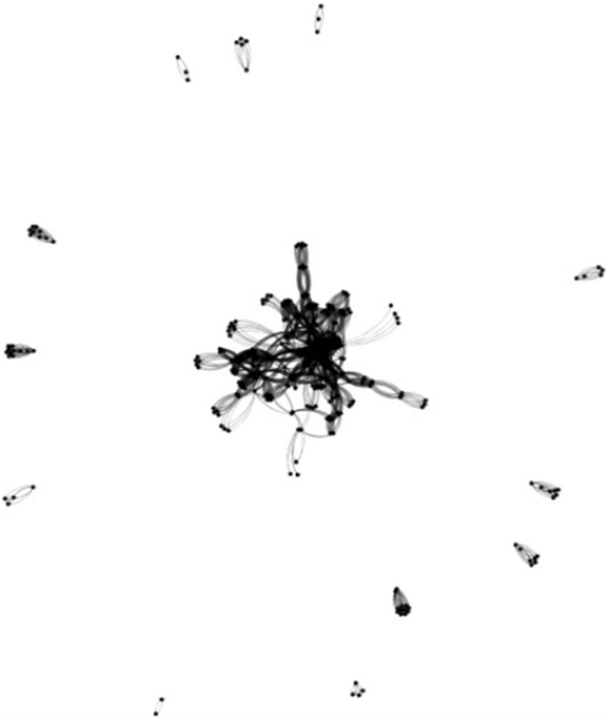

Fig. 9 Predictive network with $\tau = 0.3$ with 10,678 edges, density 0.056, average degree 24.5, and 15 connected components

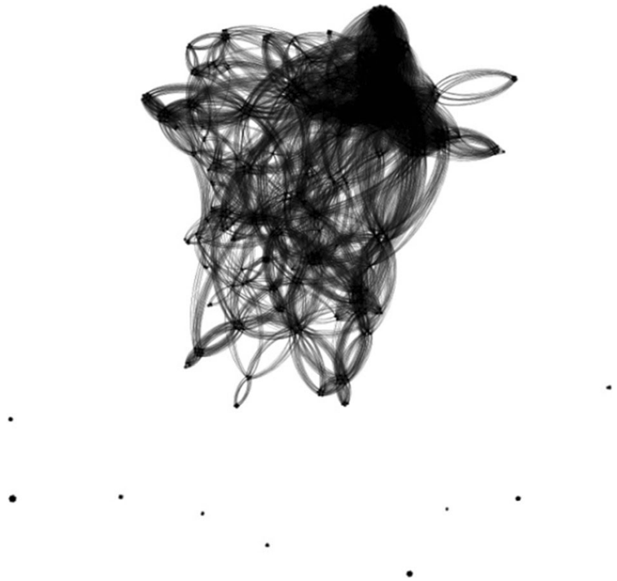

Fig. 10 Predictive network with $\tau = 0.2$ with 14,493 edges, density 0.076, average degree 33.2, and 12 connected components

3.3 Quantum communication

First, we deleted all connections with a weight of 1. The number of such edges was 376. Next, we restored links and calculated the number of restored links for different values of τ. The

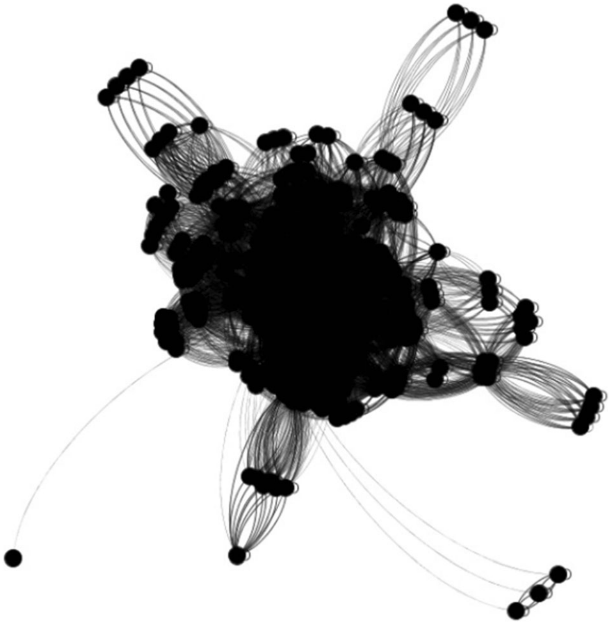

Fig. 11 Predictive network with $\tau = 0.1$ with 29,176 edges, density 0.15, average degree 66.16, and 3 connected components

Table 1 Number of restored links for various τ on the subject of quantum communication

τ	Number of restored links	Restored percentage
1.0	186	49.47%
0.9	189	50.27%
0.8	195	51.86%
0.7	222	59.04%
0.6	230	61.17%
0.5	232	61.70%
0.4	233	61.97%
0.3	253	67.29%
0.2	269	71.54%
0.1	279	74.20%

accuracy of restored links varied from 0.62 to 0.92 (with the opposite tendency relative to completeness).

The results are presented in Fig. 6–11 and Table 1; they are discussed in Section 4.

3.4 Link prediction

The same process was performed for the topic of link prediction. We deleted all links with a weight of 1 (876 links) and then restored them (and more) for various values of τ. The accuracy of restored links varied from 0.60 to 0.88. The results are presented in Fig. 12-17 and Table 2. They are discussed in Section 4.3.

4 Discussion

Figures. 6 and 12 show the original networks obtained after analysing the data for quantum communication and link prediction, respectively. These networks were then reconstructed by

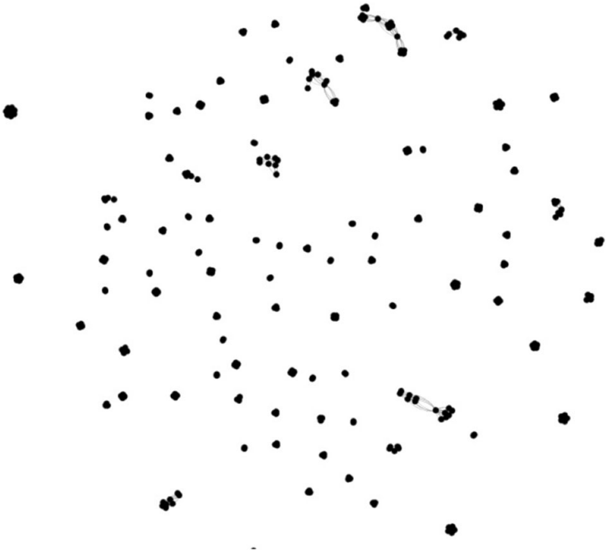

Fig. 12 Whole network – source graph with 1004 edges, density 0.009, average degree 2.99, and 89 connected components

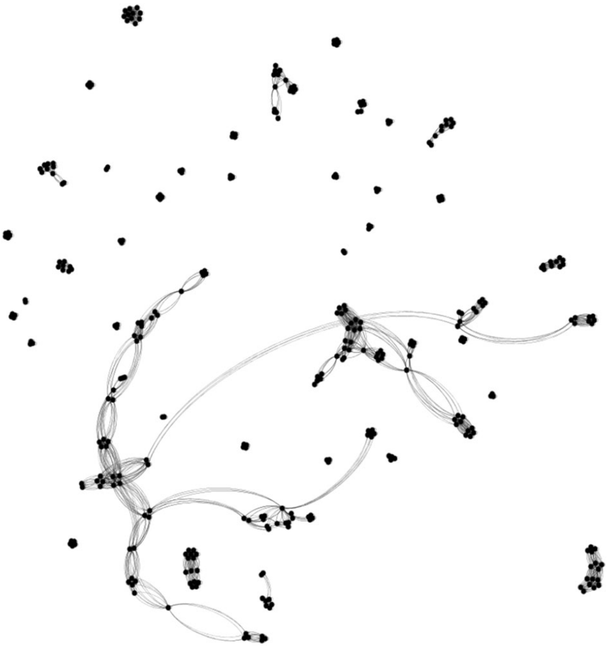

Fig. 13 Predictive network with $\tau = 0.5$ with 2382 edges, density 0.02, average degree 6.99, and 38 connected components

restoring edges with a weight $f_{a_i a_j}$ of not less than τ. The results are shown in Fig. 7–11 and 13–17. The connections with smaller weights were not taken into account, as we believe that

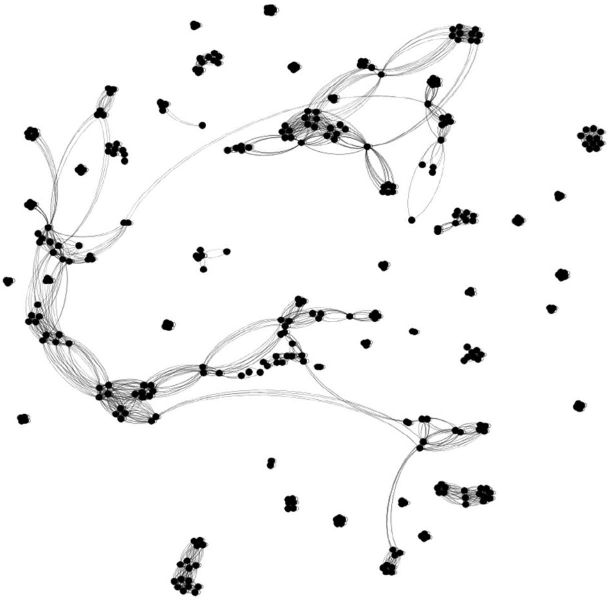

Fig. 14 Predictive network with $\tau = 0.4$ with 2484 edges, density 0.021, average degree 7.16, and 36 connected components

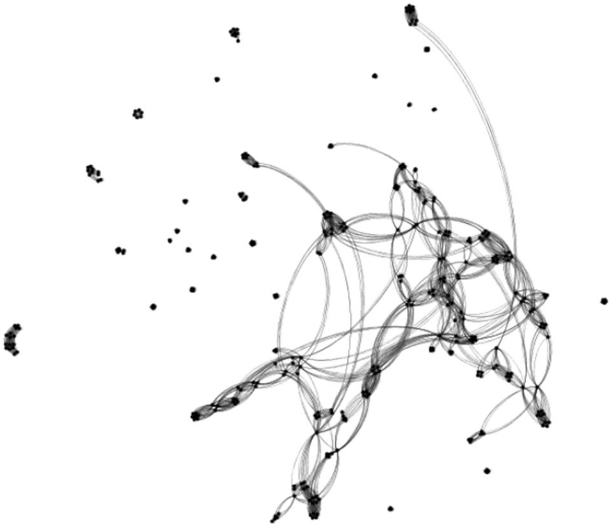

Fig. 15 Predictive network with $\tau = 0.3$ with 3154 edges, density 0.025, average degree 8.9, and 25 connected components

they do not signify a connection. The number of nodes remained constant, and the number of edges naturally increased as the value of τ decreased. The figures show an increase in network density with decreasing τ. The results presented in Tables 1 and 2 exhibit a similar trend: the number of restored links decreases as τ increases. These results demonstrate that the proposed prediction algorithm can effectively predict cooperation. In the case of link prediction, at least 78% of links are restored for any τ up to 1. However, we networks for quantum communication are relatively sparse. This sparseness of data has an impact on the prediction results.

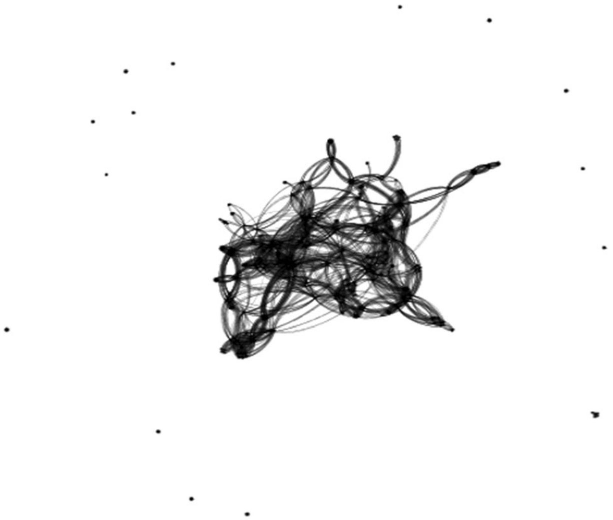

Fig. 16 Predictive network with $\tau = 0.2$ with 5232 edges, density 0.038, average degree 14.01, and 16 connected components

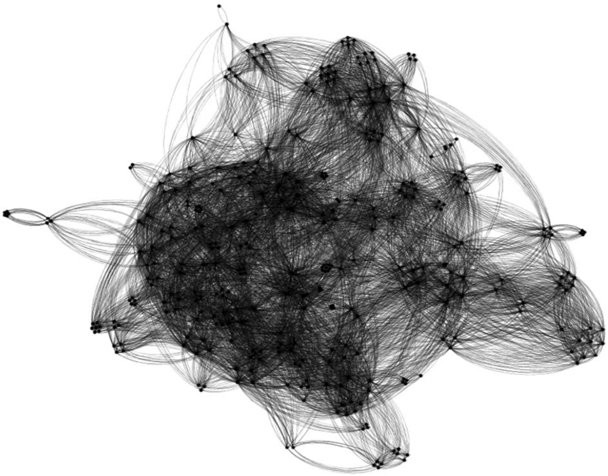

Fig. 17 Predictive network with $\tau = 0.1$ with 11,402 edges, density 0.07, average degree 28.85, and 14 connected components

Fig. 18 shows the percentage of restored links according to dates in Tables 1, 2. The experimental studies give essentially different values of the percentage of restored links; it could be explained more likely by the structure of the entire network, and not only by its density. The size of the network, the distribution of node degrees, the presence of a "rich club", etc., have a significant effect on the restoring of links. Features of network in this case are determined by the characteristics of the subject area and require further research, experiments and generalizations. At the same time, the similar behaviour of the "percentage of recovery" dependence on the threshold gives the opportunity to independently choose threshold values for researchers in the field of scientometrics within their subject areas.

The ranges of restored links in the first and second examples are different, and are largely dependent on the density of the networks. The networks considered here are realistic, taken from the Web of Science database. Even though the minimum percentage of restored links in the first example is only 49%, this is actually an excellent result—almost half of the deleted probable connections were restored based on the proposed calculation method.

Table 2 Number of restored links for various τ on the subject of link prediction

τ	Number of restored links	Restored percentage
1.0	688	78.54%
0.9	710	81.05%
0.8	747	85.27%
0.7	762	86.99%
0.6	799	91.21%
0.5	801	91.44%
0.4	814	92.92%
0.3	815	93.04%
0.2	818	93.38%
0.1	873	99.66%

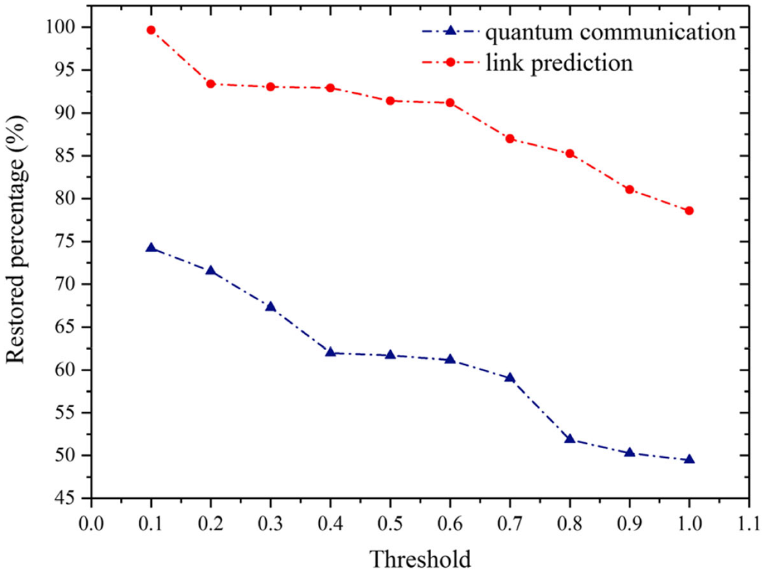

Fig. 18 Percentage of restored links for various τ on the subjects of quantum communication and link prediction

5 Conclusion

In this paper, we have proposed a new algorithm for predicting the probability of cooperation among scientists. The proposed algorithm forms a heterogeneous probabilistic network of relations between authors and descriptors from documents submitted to the Web of Science database.

On the assumption that scientific collaboration networks are heterogeneous information networks, an algorithm was developed based on meta-paths and networks with recalculated probabilistic links between nodes. Simulations of the number of restored links were conducted using real data from the Web of Science database to demonstrate the efficacy of the algorithm. Naturally, as the threshold τ increased, the number of restored links decreased and the network density increased. After removing all implicit connections (with weights of less than 1), almost 50% of the links were restored when τ was set to 1 (the worst-case scenario). The sparseness of the data had an impact on these predictive results, especially for large-degree nodes.

In future work, we plan to study the relationship between data sparsity and predictive results. Furthermore, we will investigate the use of deep learning and machine learning to predict collaboration.

Acknowledgments This work was financially supported by the Special Funding of "The Belt and Road" International Cooperation of Zhejiang Province (2015C04005).

References

1. Coccia, M., Wang, L.: Evolution and convergence of the patterns of international scientific collaboration[J]. Proceedings of the National Academy of Sciences of the United States of America. **113**(8), 2057(2016)
2. Newman, M.E.: Scientific collaboration networks. I. Network construction and fundamental results [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics. **64**(2), 016131(2001)

3. Newman, M.E.: The Structure of Scientific Collaboration Networks[J]. Proceedings of the National Academy of Sciences of the United States of America. **98**(2), 404–409(2001)
4. Newman, M.E.: Coauthorship Networks and Patterns of Scientific Collaboration[J]. Proceedings of the National Academy of Sciences of the United States of America. **101**(Suppl 1), 5200(2004)
5. Haddad, E.A., Mena-Chalco, J.P., Sidone, O.J.G.: Scholarly Collaboration in Regional Science in Developing Countries: The Case of the Brazilian REAL Network[J]. International Regional Science Review. **40**(5), 500–529(2017)
6. Abbasi, A., Hossain, L., Leydesdorff, L.: Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks[J]. Journal of Informetrics. **6**(3), 403–412(2012)
7. Milojević, S.: Principles of scientific research team formation and evolution[J]. Proceedings of the National Academy of Sciences of the United States of America. **111**(11), 3984–9(2014)
8. Kenekayoro, P., Buckley, K.: Thelwall, M.: Hyperlinks as inter-university collaboration indicators[J]. Journal of Information Science. **40**(4), 514–522(2014)
9. Peng, W., Baowen, X.U., Yurong, W.U., et al.: Link prediction in social networks: the state-of-the-art[J]. Science China Information Sciences. **58**(1), 11101–011101(2015)
10. He, Y.L., Liu, J.N.K., Hu, Y.X., et al.: OWA operator based link prediction ensemble for social network[J]. Expert Systems with Applications. **42**(1), 21–50(2015)
11. Ghasemian, F., Zamanifar, K., Ghasem-Aqaee, N., et al.: Toward a better scientific collaboration success prediction model through the feature space expansion[J]. Scientometrics. **108**(2), 777–801(2016)
12. Lu, L.Y., Pan, L., Zhou, T., et al.: Toward link predictability of complex networks[J]. Proceedings of the National Academy of Sciences of the United States of America. **112**(8), 2325–30(2015)
13. Valverde-Rebaza, J.C., Roche, M., Poncelet, P., et al.: The role of location and social strength for friendship prediction in location-based social networks[J]. Information Processing & Management. **54**(4), 475–489(2018)
14. Sett, N., Basu, S., Nandi, S., et al.: Temporal link prediction in multi-relational network[J]. World Wide Web-internet. Web Inf. Syst, **21**(2), 395–419(2018)
15. Symeonidis, P., Tiakas, E.: Transitive node similarity: predicting and recommending links in signed social networks[J]. World Wide Web-internet. Web Inf. Syst. **17**(4), 743–776(2014)
16. Gleich, D.F.: PageRank beyond the Web[J]. Computer Science. **57**, 3(2014)
17. Li, Y., Luo, P., Fan, Z.P., et al.: A utility-based link prediction method in social networks[J]. European Journal of Operational Research. **260**(2), 693–705(2016)
18. Kefalas, P., Symeonidis, P., Manolopoulos, Y.: Recommendations based on a heterogeneous spatio-temporal social network[J]. World Wide Web-internet. Web Inf. Syst. **21**(2), 345–371(2017)
19. Li, F, He, Jing, Huang, G, Zhang, Yanchun , Shi, Y and Zhou, Rui Node-coupling clustering approaches for link prediction. Knowledge-Based Systems, **89**. 669–680 (2015)
20. Muhan Zhang, Yixin Chen. Link Prediction Based on Graph Neural Networks. Advances in Neural Information Processing Systems 31, Monreal, Canada, (NIPS 2018)
21. Haeran Cho, Yi Yu. Link Prediction for Interdisciplinary Collaboration Via Co-Authorship Network, Social Network Analysis and Mining, **8** (1), 25 (2018)
22. Web of Science [DB/OL]. http://www.webofknowledge.com/wos. 21 March 2018
23. VOSviewer [EB/OL]. http://www.vosviewer.com/30 March 2018
24. Drożdż, S., Kulig, A., Kwapień, J., et al.: Hierarchical organization of H. Eugene Stanley scientific collaboration community in weighted network representation[J]. Journal of Informetrics **11**(4), 1114–1127(2017)
25. Shi, C., Li, Y., Zhang, J., et al.: A Survey of Heterogeneous Information Network Analysis[J]. IEEE Transactions on Knowledge & Data Engineering. **29**(1), 17–37(2015)
26. Gupta, M., Kumar, P., Bhasker, B.: HeteClass: a meta-path based framework for Transductive classification of objects in heterogeneous information networks[J]. Expert Syst. Appl. **68**, 106–122 (2017)
27. Chiang, M.F., Liou, J.J., Wang, J.L., et al.: Exploring heterogeneous information networks and random walk with restart for academic search[J]. Knowledge & Information Systems. **36**(1), 59–82(2013)
28. Yang, N., He, L., Li, Z., et al.: Reducing uncertainty of dynamic heterogeneous information networks: a fusing reconstructing approach[J]. Data Mining & Knowledge Discovery. **31**(3), 879–906(2017)
29. Ma, Y., Yang, N., Zhang, L., et al.: Predicting neighbor label distributions in dynamic heterogeneous information networks[J]. World Wide Web-internet. Web Inf. Syst. **20**(6), 1269–1291(2017)
30. Ozcan, A., Oguducu, S.G.: Link prediction in evolving heterogeneous networks using the NARX neural networks[J]. Knowledge & Information Systems. **55**(2), 333–360(2018)
31. Li, J.C., Zhao, D.L., Ge, B.F., et al.: A link prediction method for heterogeneous networks based on BP neural network[J]. Physica A Statistical Mechanics & Its Applications. **495**, 1–17 (2018)
32. Sun, Y., Han, J.: Meta-Path-Based Search and Mining in Heterogeneous Information Networks[J]. Tsinghua Science and Technology. **18**(4), 329–338(2013)

33. Fu, G., Ding, Y., Seal, A., et al.: Predicting drug target interactions using meta-path-based semantic network analysis[J]. Bmc Bioinformatics. **17**(1), 160(2016)
34. Qian, F., Gao, Y., Zhao, S., et al.: Combining topological properties and strong ties for link prediction[J]. Tsinghua Science and Technology. **22**(6), 595–608(2017)
35. Huang, R., Yang, H., Bei, S., et al.: Bioinformatic Analysis Identifies Three Potentially Key Differentially Expressed Genes in Peripheral Blood Mononuclear Cells of Patients with Takayasu's Arteritis[J]. Cell Journal. **19**(4), 647–653(2018)
36. Liu, Y., Zeng, X., He, Z., et al.: Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics. pp.1–1 (2016)
37. Gephi[EB/OL]. https://gephi.org/30 March 2018
38. Newman. M.E.: The Structure and Function of Complex Networks [J]. Siam Review. **45**(2), 167–256(2003)

Affiliations

Dmytro Lande [2,3] **· Minglei Fu** [1] **· Wen Guo** [1] **· Iryna Balagura** [2] **· Ivan Gorbov** [2] **· Hongbo Yang** [1]

✉ Minglei Fu
fuml@zjut.edu.cn

[1] College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

[2] Institute for Information Recording, NAS of Ukraine, Kiev 03113, Ukraine

[3] Information Research Institute, Shandong Academy of Sciences, Jinan 250014, China