International Journal of Modern Physics C
Vol. 28, No. 6 (2017) 1750082 (12 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0129183117500826

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

Clustering-based link prediction in scientific coauthorship networks

Yang Ma*, Guangquan Cheng†, Zhong Liu‡
and Xingxing Liang§

Science and Technology on Information
Systems Engineering Laboratory
National University of Defense Technology
Changsha 410073, P. R. China
**yang_ma_cn@163.com*
†cgq299@163.com
‡philipliu@163.com
§doublestar_l@163.com

Received 22 April 2016
Accepted 25 April 2017
Published 26 May 2017

Link prediction in social networks has become a growing concern among researchers. In this paper, the clustering method was used to exploit the grouping tendency of nodes, and a clustering index (CI) was proposed to predict potential links with characteristics of scientific cooperation network taken into consideration. Results showed that CI performed better than the traditional indices for scientific coauthorship networks by compensating for their disadvantages. Compared with traditional algorithms, this method for a specific type of network can better reflect the features of the network and achieve more accurate predictions.

Keywords: Complex networks; spectral clustering; link prediction; scientific coauthorship.

PACS Nos.: 11.25.Hf, 123.1K.

1. Introduction

In recent years, link prediction in social networks has become of interest in the research community. Link prediction aims to infer the probability of establishing a link between two nodes based on observed links and properties of the nodes.[1,2] There are two primary aims: (i) Predict unknown links. For most real networks, not all links are observable. Link prediction helps identify the missing links. (ii) Predict future links. Based on the present network, link prediction helps infer the possibility of a link in a future network.

Link prediction is an important aspect of social network analysis, which is significant both in theory and in practice.[3] Theoretically, research on link prediction helps people understand the evolution mechanism of social networks. Practically, it is

†Corresponding author.

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

widely used in scientific applications. In protein molecular networks, link prediction technology improves the efficiency compared with traditional molecular detection by experiment and therefore lowers the cost. In social networks, possible friends and interesting topics are recommended to the users by link prediction technology, which improves the user experience. In recent years, scientific research cooperation networks have become popular. Researchers in various scientific fields have found that these networks are scale-free, which is a major conceptual breakthrough. Scientific collaboration occurs when researchers find collaborators from different scientific areas and professions to jointly complete academic research. Using link prediction technology, potential collaborators are recommended to researchers to achieve unique scientific discoveries.

Two types of link prediction methods have been developed[4]: node-property-based and network-structure-based. The properties of the node are usually invisible or inaccessible, making it hard to utilize node-property-based methods. However, node connections are accessible, making network-structure-based methods attractive. Commonly used structure-based methods include common-neighbor-based methods, node-degree-based methods, and path-based methods. These methods begin with the structure of the network and use network topology information to predict possible future edges through computing structure similarity. This paper studies the link prediction problem using a network-structure-based method.

The link prediction methods described above produce good predictions, but two issues remain unsettled.[5,6] First, clustering is a common and basic physical phenomenon in networks, but is not accurately modeled with link prediction methods. Even though most network indices reflect grouping, it is not known how clustering will influence the predicted results. Second, improvements for specific type of networks are not applied. Networks of different types reach different prediction accuracy using the same index.

In this paper, an innovative clustering-based method is proposed to solve the two problems mentioned above. The clustering characteristics of nodes are first examined to find potential links. Nodes are separated into different clusters by the clustering algorithm. Some nodes are close to the cluster center, indicating they exhibit obvious cluster features, while others are next to the cluster edge, indicating they exhibit indistinct cluster features. The clusters and positions of the nodes within the clusters determine the possibility of a link between two nodes. Additionally, the algorithm is modified to study coauthorship networks to achieve more accurate predictions. Through the analysis of correlation properties between nodes in scientific coauthorship networks, characteristics of this type of network are found and a targeted algorithm is designed to enhance the accuracy of the link predictions.

The structure of the paper is organized as follows. In Sec. 2, the network is defined and a clustering index is proposed. Section 3 outlines the overall processing of data. Traditional algorithms are provided for comparison in Sec. 4. Several network simulations are illustrated in Sec. 5, along with their results. Conclusions and future work are then provided in Sec. 6.

2. Algorithm Description

The purpose of clustering algorithms is to quantify the positions of the nodes in the network after clustering. Spectral clustering[7,8] is a classic method used in complex networks, which treats network clustering as a quadratic optimization problem. Eigenvalues and eigenvectors are calculated to build a simplified data space. The NJW algorithm[9] proposed by Ng *et al.* is one example. The NJW algorithm uses similar matrices to cluster eigenvectors. Eigenvectors corresponding to the n largest eigenvalues are selected from the structure matrix, and then in the n-dimensional space, k-means or other simple algorithms are used for clustering. The characteristics of the spatial distribution of each node with a set of coordinate annotations allows for a quantitative index to be determined. The multidimensional coordinate method can be used to cluster nodes quickly and obtain quantitative multidimensional geometric centers of each cluster. By computing the Euclidean distance between a node and the geometric center of the cluster to which it belongs, the position of the node in the class can be measured. The smaller the distance, the closer to the cluster center this node is.

The coauthorship network is defined as $G = (V, E)$ with sets of nodes and edges denoted by V and E, respectively. Links in G are denoted by the adjacent matrix A, in which (v_i, v_j) is a nonnegative integer representing the link of node i and j. $n = |V| = \text{rank}(G)$ is the number of nodes in the network. In the adjacency matrix, one row or one column represents the connection between one node and the other nodes in the network with magnitudes corresponding to the weight of the links. For a scientific coauthorship network, each row or column represents a researcher and each value in the matrix indicates the magnitude of cooperation between two researchers. The adjacency matrix is the input to the spectral clustering algorithm.

The number of clusters N controls the clustering quality. A good balance between compression and accuracy is the optimal outcome. If the entire dataset is represented by a single cluster, the data compression is maximized, but the clustering has no value. On the other hand, if each data point is represented as a cluster, it will produce the finest clustering (i.e. the most accurate solution, because the distance of a node to its corresponding cluster center is zero), which similarly cannot provide any meaningful information. There are two typical methods to determine N. The experience-based[10] method to determine $N = \sqrt{n/2}$ is useful in social networks. However, this method can only provide a rough number of clusters. The elbow method[10] identifies the first turning point that minimizes the sum of intra-variance in the clusters. The elbow method is based on the following phenomena: increasing the number of clusters reduces the variance because more precise data clustering can be achieved; however, if too many clusters are formed, the variance cannot accurately reflect the network behavior. Therefore, the correct heuristic method is to use the first turning point over the curve of sum of variance as a function of the number of clusters in the network. The elbow method performs well in collaboration networks. Therefore, the second method is adopted in this paper.

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

In scientific coauthorship networks, researchers are divided into different categories according to their research fields. Researchers close to cluster centers concentrate more on specific subjects. As a result, their cooperation relationships tend to occur in the same field. For researchers closer to the cluster edge, their cooperative relationships occur both in and out of their research field, so there is more cross-subject collaboration. Partnerships between researchers are mainly in two forms:

- in the same field (cluster), researchers closer to the edge of the field tend to cooperate with partners closer to the center to increase the influence of the research;
- in different fields (clusters), researchers on the edge of one field work with those closer to the center of another cluster to supplement their lack of knowledge in this field.

In a word, researchers closer to the edge of fields are more inclined to cooperate with researchers in the center.

Using the spectral clustering method, each node is characterized by an eigenvector, which reduces the computational dimension. Eigenvectors of the matrix form the eigenmatrix $X^{n\times N}$, which is normalized to $Y^{n\times N}$. Using k-means on Y, we can get the node grouping. Each eigenvector in the eigenmatrix can be regarded as the coordinate of the node in n-dimensional space from the point of view of the other nodes. The Euclidean distance of the node to the center of the cluster is selected as the link index, which determines the possibility of generating links in a network. dist satisfies

$$\mathrm{dist}_x = \|Y_x - c_i\|, \quad x \in C_i, \tag{1}$$

where Y_x denotes the coordinate of node x in matrix Y in eigenvector form. C_i denotes the cluster that node x belongs to and $|C_i|$ represents the number of nodes in cluster C_i. c_i denotes the coordinates of the center of cluster i such that

$$c_i = \frac{1}{|C_i|}\sum_{k\in C_i} Y_k. \tag{2}$$

To compare different clusters, we normalize dist by the largest value in the network and get

$$s\,\mathrm{dist}_x = \mathrm{dist}_x / \max\nolimits_{k\in V}\mathrm{dist}_k. \tag{3}$$

Based on the physical phenomena in scientific coauthorship networks, the clustering index (CI) is given below

$$\mathrm{CI}_{xy} = \sin\frac{\pi}{2}(s\,\mathrm{dist}_x + s\,\mathrm{dist}_y). \tag{4}$$

Since the value of $s$dist is normalized to $[0,1]$ and the range of the sine function is limited to $[0,\pi]$, the index reaches a higher value when the sum of the two nodes'

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

Algorithm 1 Clustering index based on spectral clustering

```
• Input
    (a) $A$: adjacency matrix
    (b) $n = \mathrm{rank}(A)$: number of nodes
• Output
    (a) CI: clustering index matrix
• Procedure
    (a) Determine $N$: elbow method
    (b) Define $D$ diagonal matrix, whose $(i, i)$ is the sum of $A$'s $i$th row, and define $L = D^{-1/2}AD^{-1/2}$
    (c) Find the largest $N$ eigenvectors $x_1, x_2, ..., x_N$ of $L$ (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $X = [x_1 x_2 ... x_N] \in \mathbb{R}^{n \times N}$
    (d) Form matrix $Y^{n \times N}$ by normalizing each row of $X$ to have unit length
    (e) Cluster ($k$-means) rows of $Y$ into $N$ disjoint clusters denoted as $C$
    (f) Calculate each class center $c$ in $Y$ with $c_i = \sum_{k \in C_i} Y_k / |C_i|$
    (g) Calculate nodes' distance to their own class center dist and normalize it to $s$dist with $\mathrm{dist}_x = \| Y_x - c_i \|, x \in C_i$ and $s\mathrm{dist}_x = \mathrm{dist}_x / \max_{k \in V} \mathrm{dist}_k$
    (h) Calculate clustering index $\mathrm{CI}_{xy} = \sin \frac{\pi}{2}(s\mathrm{dist}_x + s\mathrm{dist}_y)$
    (i) Return CI
```

$s$dist is closer to 1, which is consistent with the two mentioned phenomena. The CI computation algorithm is described in Algorithm 1.

3. Data Processing

Cooperation between high-energy physics researchers (ca-cit-HepTh[11]) is used as input data to test the clustering algorithm. In this section, the original data are processed for better algorithm performance. The data structure is $Data[from, to, weight, timestamp]$. Each record is composed of four parts: *from*, *to*, *weight* and *timestamp*. The first two attributes define the edge of the network, while the third and fourth attributes indicate its weight and time information, respectively. To form a network, each record is transformed to a link marked by unique weight and time information.

This network covers 28 093 nodes and 4 596 803 relations. To decrease uncertainty influenced by the small number of nodes and for better comparison, the maximal connected subgraph of the most active nodes in the network are selected as

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

```
Algorithm 2 Data processing
• Input
    (a) data: original network
• Output
    (a) $A$: adjacency matrix
• Procedure
    (a) Calculate total degrees $degree[n]$ of every node
    (b) Find $x$ largest nodes in $degree[n]$ to form $select_1[x]$ (undirected and unweighted) and corresponding adjacency matrix $local_review$, bubble sort nodes in $select_1$
    (c) Find maximal connected subgraph in $select_1$ listed as $select_2[y]$
    (d) Select records of $select_2$ from $data$ and transfer $timestamp$ into seasonal information $relevant_time$
    (e) Transfer $data$ into adjacency matrix $A$ according to $relevant_time$
    (f) Return $A$
```

input data, which is divided into 15 matrices by season according to timestamp. Detailed data processing is described in Algorithm 2.

Each row or column in the adjacency matrix shows the cooperative relationships of one researcher. The numerical value in the matrix indicates the number of cooperative papers in the network (undirected). The processed networks consist of 114 nodes and 2361 cooperative relationships. The number of specific cooperation relationships in each network are listed in Table 1.

The effect of the number of clusters on the sum of variance for the coauthorship network is shown in Fig. 1. According to the elbow method, $N = 3$ is selected for

Table 1. Detailed cooperation relationships in each network.

Group	Coauthorships	Group	Coauthorships
1	145	9	74
2	210	10	90
3	228	11	142
4	60	12	160
5	283	13	87
6	206	14	200
7	165	15	95
8	216		

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

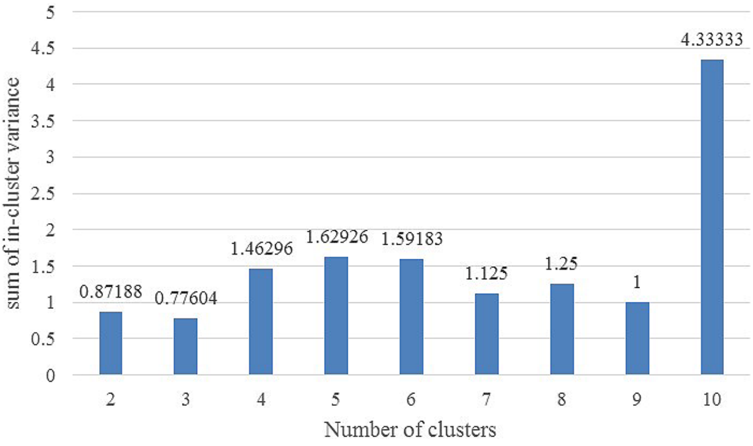


Fig. 1. (Color online) Sum of in-cluster variance on number of clusters.

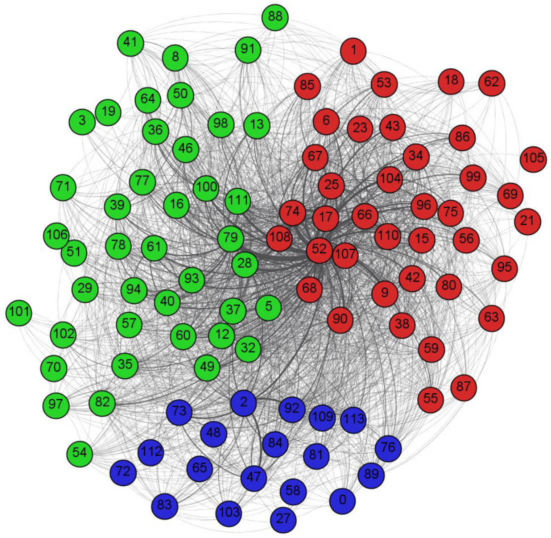


Fig. 2. (Color online) Global view of processed network.

the simulations in this paper, which means researchers are divided into three distinct groups. The corresponding global network after grouping is shown in Fig. 2, with each color representing one group.

4. Related Algorithms

To quantify the performance of the algorithm, we compare the results with common-neighbor-based and node-degree-based link prediction methods using the same

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

dataset. Common neighbor (CN) is a representative example of common-neighbor-based link prediction methods. In CN, the more common neighbors two nodes have, the higher degree of similarity is assigned.

$$\mathrm{CN}_{xy} = |\Gamma(x) \cap \Gamma(y)|, \tag{5}$$

where CN_{xy} represents the index between node x and y and $\Gamma(x)$ is the neighbor set of node x. In the weighted network, the CN formula is modified to calculate the sum of weights of common neighbors.

$$\mathrm{WCN}_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} (w_{xy} + w_{yz}), \tag{6}$$

where z represents the common neighbors and w_{xz} is the weight between x and z.

Similar methods are the Jaccard[12] index, Salton[13] index and HPI[14] index, shown in Table 2. These indicators use the same basic calculation, but differ in measuring the degrees of nodes.

A representative method of degree-based link prediction is the adamic/adar (AA)[15] index, in which common neighbor nodes with small degrees have greater influence than those with larger degrees. Small degree nodes are given higher weights so that they are more likely to affect the generation of links.

$$\mathrm{AA}_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log d(z)}. \tag{7}$$

A similar index is resource allocation (RA)[16] from network RA. For two unconnected nodes x and y in a network, if x transfers some resources to y, the transfer process is dependent on the common neighbors. Considering the simplest case, one unit of resources of node x will be transferred to all its neighbors, then the similarity between the nodes can be expressed by the amount of resources received from node x.

$$\mathrm{RA}_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{d(z)}. \tag{8}$$

AA and RA have similar mathematical expressions, both weakening the degree of contribution of the node in common neighbors. The difference is that the AA index is

Table 2. Common neighbor-based link prediction methods.

Index	Formula
Jaccard	$\mathrm{Jaccard}_{xy} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Salton	$\mathrm{Salton}_{xy} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\sqrt{d(x)\times d(y)}}$
	$d(x)$ is degree is of node x
HPI	$\mathrm{HPI}_{xy} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\min d(x),d(y)}$

a logarithmic function of the degree while the RA index is a linear function of the degree. The difference between the two is not obvious when the degrees of common neighbors are relatively small, while the difference between the two is obvious in large degree cases. That is, the RA index weakens the influence of large degree nodes more than the AA index. In networks with large degrees, the performance of RA is superior to AA. In weighted networks, the total weights of common neighbors are taken into consideration.

$$\mathrm{WRA}_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{yz}}{d(z)}. \tag{9}$$

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

5. Results

Area under ROC curve (AUC)[17] is commonly used to judge the performance of link prediction algorithms. One existent link and another inexistent link are chosen randomly from the testing data, and if the former scores higher than the latter, add one; if the two score equally, add 0.5. After repeating n times, where n' times satisfies the former condition and n'' satisfies the latter, the AUC index is given by

$$\mathrm{AUC} = \frac{n' + 0.5n''}{n}. \tag{10}$$

According to this formula, if all the edges are randomly generated, AUC is equal to 0.5. The link prediction algorithm is more accurate if the AUC index value is closer to one.

For 15 consecutive quarters of data, with any quarterly data as input and the following as test, the AUC values of the CI index along with the CN, WCN, RA and WRA indices are calculated for comparative analysis. A total of 14 comparative tests are carried out with each index repeated 1000 times to calculate AUC. For accuracy, CI produces the best clustering results from 10 repetitions.

AUC reflects the global accuracy of link prediction algorithms. By comparing the 14 groups of results in Table 3, we find that in scientific coauthorship networks, CI performs much better than CN, WCN, RA and WRA, which shows high global prediction ability of CI over the other indices.

AUC is a straightforward way to judge the global performance of algorithms, but for practical requirements, finding possible links is of higher importance. Another metric is Precision,[6] which quantifies the prediction accuracy of the largest L edges. If m edges are correctly predicted, the precision is defined as

$$\mathrm{Precision} = \frac{m}{L}. \tag{11}$$

The higher the Precision, the more accurate the prediction. For two prediction results with the same AUC, the one with higher Precision performs better, because it tends to prioritize correct links. As a result, Precision usually works as an assistant for AUC.

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

Table 3. AUC of CN, WCN, RA,WRA and CI.

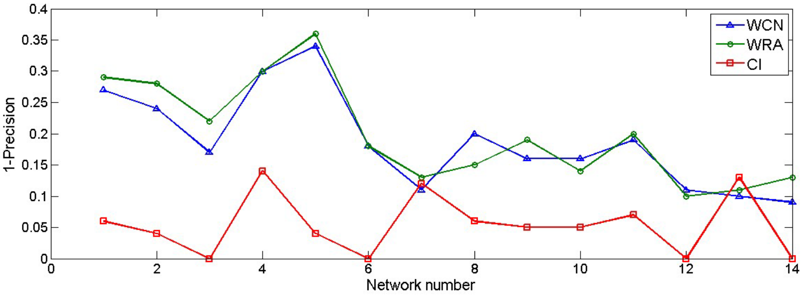
Group	CN	WCN	RA	WRA	CI
1	0.674	0.6585	0.6825	0.673	**0.775**
2	0.7405	0.7405	0.7375	0.7435	**0.836**
3	0.7935	0.809	0.807	0.802	**0.82**
4	0.6405	0.651	0.643	0.628	**0.767**
5	0.7115	0.6825	0.7205	0.704	**0.73**
6	0.6775	0.694	0.68	0.685	**0.802**
7	0.703	0.6915	0.6745	0.6925	**0.759**
8	0.678	0.693	0.6785	0.7005	**0.784**
9	0.6265	0.622	0.6445	0.619	**0.7805**
10	0.574	0.5805	0.5495	0.572	**0.789**
11	0.645	0.6455	0.6405	0.635	**0.718**
12	0.694	0.6875	0.682	0.7025	**0.76**
13	0.5775	0.578	0.5745	0.593	**0.680**
14	0.747	0.757	0.7395	0.768	**0.775**

For the Precision index, determining the value of L is crucial. In this paper, several groups of data are utilized. It is inappropriate to adopt the same L in different networks. k-Precision is used to satisfy this need. Denoting the search range as k and the total number of future links as M, the proportion of correctly predicted links m in the largest kM edges is given by

$$k - \text{Precision} = \frac{m}{M}. \tag{12}$$

We calculate the k-Precision of WRA, WCN and CI with $k = 1, 3, 5, 7$ in 14 networks separately. Results are shown in Fig. 3.

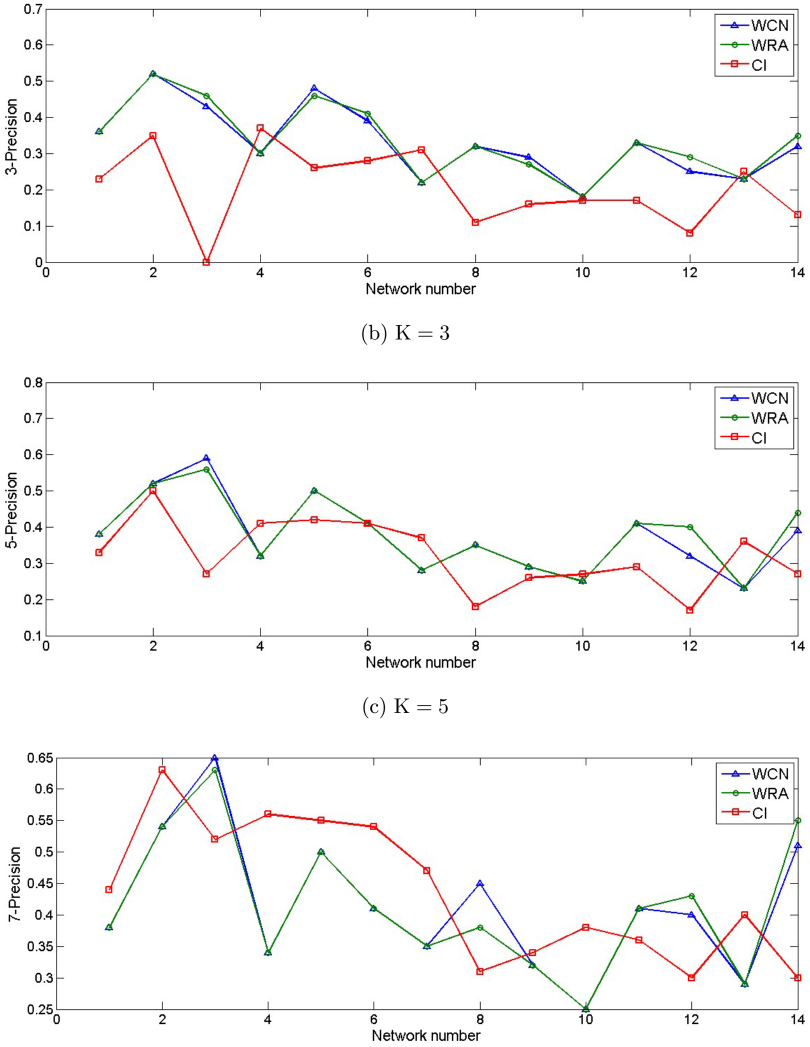
k-Precision focuses on the accuracy of links with higher possibility. WCN and WRA achieve similar outcomes according to this index, managing to accurately find possible links in the smaller range (smaller k) while performing poorly in the larger range (higher k). CI performs poorly in the smaller range but covers more links in the



(a) K = 1

Fig. 3. (Color online) k-Precision with $k = 1, 3, 5, 7$.

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.



(b) K = 3

(c) K = 5

(d) K = 7

Fig. 5. (*Continued*)

larger range, which is consistent with the AUC analysis. In summary, WRA, WCN, and similar algorithms aim to find links with high prediction value, but perform poorly from a global standpoint. The opposite is true for CI, which finds more possible links but ignores the highest ones. Combining these algorithms will achieve better prediction results.

6. Conclusion

In this paper, the spectral clustering algorithm is applied to grouping nodes and characteristics of scientific coauthorship networks, and CI is developed for link prediction. A new judgment index k-Precision is proposed to better compare and analyze calculating results. Compared with traditional indices using 14 groups of input data, the network-based index more accurately reflects the essence of the network and performs better from a global standpoint. CI is a better predictor of potential links, which compensates for the disadvantages of the traditional indices.

Acknowledgments

This paper is supported in part by the National Natural Science Foundation of China under Grant Nos. 61201328 and 71471175, Doctoral Fund of Ministry of Education of China under Grant No. 20114307120021 and Hunan Provincial Natural Science Foundation of China under Grant No. 13JJ4011.

Int. J. Mod. Phys. C Downloaded from www.worldscientific.com
by THE UNIVERSITY OF HONG KONG on 06/13/17. For personal use only.

References

1. L. Lin-Yuan, *J. Univ. Electron. Sci. Technol. China* **39**, 651 (2010).
2. L. Linyuan and T. Zhou, *Physica A, Stat. Mech. Appl.* **390**, 1150 (2010).
3. L. Wang and X. Li, *Sci. Bull.* **59**, 3511 (2014).
4. Y. Ma, G. Cheng, Z. Liu, F. Xie, K. A. Dawson, J. O. Indekeu, H. E. Stanley and C. Tsallis, *Physica A, Stat. Mech. Appl.* **465**, 792 (2016).
5. J. Kim, M. Choy, D. Kim and U. Kang, Link prediction based on generalized cluster information, in *Companion Publication of the Int. Conf. World Wide Web Companion* (2014).
6. J. C. Valverde-Rebaza and A. D. A. Lopes, *Lect. Notes Comput. Sci.* **4**, 92 (2012).
7. P. Symeonidis and N. Mantas, *Soc. Netw. Anal. Min.* **3**, 1433 (2013).
8. X. Feng, J. C. Zhao and K. Xu, *Phys. Condens. Matter* **85**, 1 (2011).
9. A. Y. Ng, M. I. Jordan and Y. Weiss, *Proc. Adv. Neural Inf. Process. Syst.* **14**, 849 (2001).
10. I. H. Witten and E. Frank, *Biomed. Eng. Online* **5**, 95C97 (2011).
11. J. Leskovec, J. Kleinberg and C. Faloutsos, *ACM Trans. Knowl. Discov. Data* **1**, 2 (2006).
12. A. Yong-Yeol, J. P. Bagrow and L. Sune, *Nature* **466**, 761 (2010).
13. F. Q. Yang, T. L. Sun and J. G. Sun, *Wuhan Univ. J. Nat. Sci.* **11**, 6 (2006).
14. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edn. (Morgan Kaufmann Publishers Inc., 2005).
15. M. O. Jackson and A. Wolinsky, *J. Econ. Theory* **71**, 44 (1996).
16. T. Zhou, L. Linyuan and Y. C. Zhang, *Phys. Condens. Matter* **71**, 623 (2009).
17. N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective* (Cambridge University Press, 2011), pp. 100–110.