Hindawi
Complexity
Volume 2021, Article ID 8889441, 5 pages
https://doi.org/10.1155/2021/8889441

Research Article

A Method for Improving the Accuracy of Link Prediction Algorithms

Jie Li[1], **Xiyang Peng**[1], **Jian Wang**[2] and **Na Zhao**[1,3]

[1]*Key Laboratory in Software Engineering of Yunnan Province, School of Software, Yunnan University, Kunming 650091, China*
[2]*College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China*
[3]*Electric Power Research Institute of Yunnan Power Grid, Kunming 650217, China*

Correspondence should be addressed to Na Zhao; zhaona@ynu.edu.cn

Received 7 September 2020; Revised 12 April 2021; Accepted 10 May 2021; Published 24 May 2021

Academic Editor: Anirban Chakraborti

Link prediction is a key tool for studying the structure and evolution mechanism of complex networks. Recommending new friend relationships through accurate link prediction is one of the important factors in the evolution, development, and popularization of social networks. At present, scholars have proposed many link prediction algorithms based on the similarity of local information and random walks. These algorithms help identify actual missing and false links in various networks. However, the prediction results significantly differ in networks with various structures, and the prediction accuracy is low. This study proposes a method for improving the accuracy of link prediction. Before link prediction, *k*-shell decomposition method is used to layer the network, and the nodes that are in 1-shell and the nodes that are not linked to the high-shell in the 2-shell are deleted. The experiments on four real network datasets verify the effectiveness of the proposed method.

1. Introduction

Link prediction refers to the prediction of the possibility of a link between two nodes in the network that have not yet formed a link or inference of the missing link in the network through the known network topology, network node attributes, and other information [1]. Link prediction can be used in different fields because this process predicts not only the lost links in the network [2] but also the possible future links in the evolved network. In social networks, recommending links that do not yet exist as potential friendship relationships can help users find new friends [3]. In the battlefield communication networks, link prediction supplements the network topology reconnaissance by predicting the hidden links and the links that are not yet detected by the enemy's communication network to provide conditions for network confrontation [4]. In biological networks, such as protein-protein interaction and metabolic networks [5], a large number of experiments is required to prove that the two nodes in these networks are connected. However, completing these experiments is costly and time-consuming. Therefore, predicting the unknown links in the network based on the known ones is more reasonable and effective, checking all potential links without solid basis. Link prediction is also often applied to personalized recommendations. For example, in e-commerce systems, predictions are made based on user clusters and purchase relationships first before providing personalized product recommendations to different users to improve the platform's performance and order quantity [6]. In summary, link prediction research has important theoretical and practical application values. In addition, the accuracy of link prediction algorithms can be further improved when applied to real networks, especially in social networks. In this regard, this study proposes a method for improving the accuracy of link prediction algorithms.

Numerous link prediction algorithms exist. However, scholars in various fields are particularly interested in the similarity method based on network structure, which has low complexity and good effect. Common neighbor (CN) [7], which is the simplest similarity index, describes similarity by calculating the number of common neighbors

between any two nodes. On the basis of CN, researchers proposed other local similarity indicators, such as Jaccard Index [8], Adamic and Adar [9], and resource allocation (RA) indices [10], which demonstrate better prediction effect than the former. Barabási and Albert [11] proposed the preferential attachment process based on degree. Lü et al. [12] considered the third-order path on the basis of CN and proposed a local path, which produces excellent results at the expense of a certain degree of complexity. Liu et al. [13] proposed the extended resource allocation (ERA) algorithm on the basis of RA. Li et al. [14] introduced the potential information capacity (PIC) index. Some of the above-mentioned indicators, such as RA, ERA, and PIC, portray and apply different perspectives to the resource transmission process in complex networks. In consideration of the global information of the network, global indicators (e.g., full path indicator algorithm KATZ [15], average commute time [16], and cosine similarity indicator Cos+ [17]), which obtain better effects than local methods, are also proposed. However, the complexity of such indicators is high, and applying them to large and complex networks is difficult. RA achieves excellent results in the case of low-complexity networks, which are close or even higher than the global index in some networks. Kovács et al. [18] proposed the L3 algorithm, which produces satisfactory results in protein interaction networks. With the advancement of the study on network evolution mode and link generation mechanism, the accuracy of link prediction algorithms can be further improved.

The main contributions of this study are as follows:

We infer that in the k-shell decomposition algorithm, if the nodes in the outer shell share links with those in the high-shell, the possibility of linkage between the former and other nodes will increase.

We propose a method for improving the accuracy of the link prediction algorithm by deleting the outer nodes of the k-shell. This approach provides insights into the enhancement of the link prediction accuracy.

We evaluate the proposed method on four different network datasets, compare state-of-art algorithms [8, 16] on different percentages of test sets, and verify the ability of the proposed method to improve the link prediction accuracy.

2. Methods

The existence of a link between any two nodes in the network can be speculated via link prediction. However, some links may not exist in the actual network; such a phenomenon will reduce the accuracy of link prediction. Deleting nodes with low potential link probability before link prediction might improve the accuracy of link prediction. On this basis, we design the following method:

First, the 1-shell is obtained in accordance with the k-shell method [19]. All nodes with a degree of 1 and their corresponding edges are removed. If news nodes with a degree of 1 appear in the network, these nodes and their connected edges are removed. This operation is repeated until no nodes with a degree of 1 are present in the network. The removed nodes and the connected edges between them are the 1-shell of the network. The 2-shell, 3-shell... are obtained using the same procedure.

The node located in the 1-shell denotes the person who is not good at socializing in the social network. The possibility of this person making new friends is extremely low. Therefore, we delete the nodes that are in the 1-shell. If the node in the 2-shell is not connected to a node that is in a high-shell, it is also deleted. As shown in Figure 1, nodes 9 and 11 are in 1-shell, so they are deleted. Nodes 12, 13, 14, and 15 are in the 2-shell. Since nodes 13 and 14 are not connected to x-shell ($x \geq 4$) nodes, they are deleted. Figure 2 is what we get at this point. We then use the link prediction algorithm. If the k-shell contains many shells or k is large, the nodes in the 1-shell and 2-shell can be deleted and nodes in 3-shell that are not connected to x-shell ($x > 4$) or more outer shells can be deleted.

3. Experiments and Results

We conducted experiments on four real network datasets in different fields. Given that this study considers undirected and unweighted networks, the weight and direction of the edges in the network are disregarded. These details of the four networks are discussed in the following paragraphs:

(i) Lesmis: it is a network of the relationship among the characters of the film "Les Misérables." A node represents a character and an edge between two nodes shows that these two characters appeared in the same chapter of the book [20].

(ii) Polbooks: it is a network of the purchases of the United States political books [21].

(iii) Metabolic: this is the metabolic network of the roundworm *Caenorhabditis elegans*. Nodes are metabolites (e.g., proteins), and edges are interactions between them. Since a metabolite can iterate with itself, the network contains loops. The interactions are undirected. There may be multiple interactions between any two metabolites [22].

(iv) Net-science: this is a network of coauthorships in the area of network science [23].

Table 1 summarizes the basic topology features of these networks, where N, M, k, and C represent the number of network nodes, number of network edges, average degree, and clustering coefficient, respectively.

Area under the curve (AUC) indicators [24] are commonly used to evaluate the accuracy of link prediction. The network is divided into a training set and a test set. After calculating the link score between every two nodes in the network through the training set, each time of an edge is randomly selected from the test and nonexistent edge sets for comparison. If the similarity score value of the edge in the test set is greater than that in the nonexistent edge set, one point is added (n_2 times in total). If the two similarity score values are equal, 0.5 point is added (n_1 times in total). Considering independent comparison for n times, the AUC is expressed as

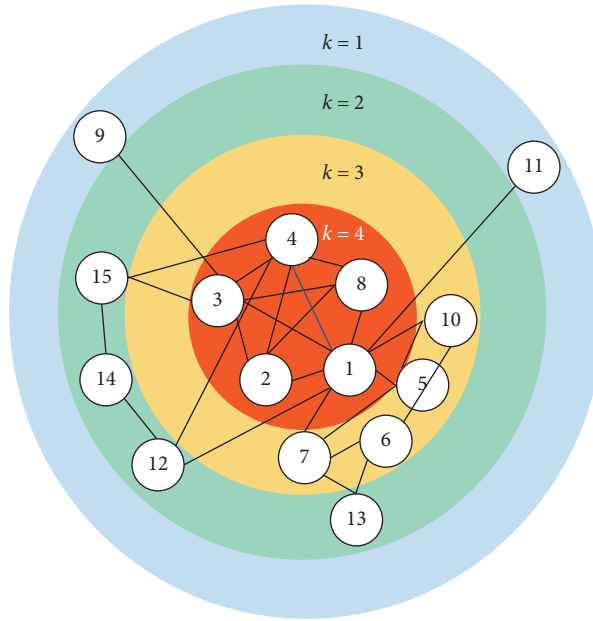

Figure 1: A k-shell layered diagram. Nodes 1, 2, 3, 4, and 8 are in the 4-shell; nodes 5, 6, 7, and 10 are in the 3-shell; nodes 12, 13, 14, and 15 are in the 2-shell, and the rest are in the 1-shell.

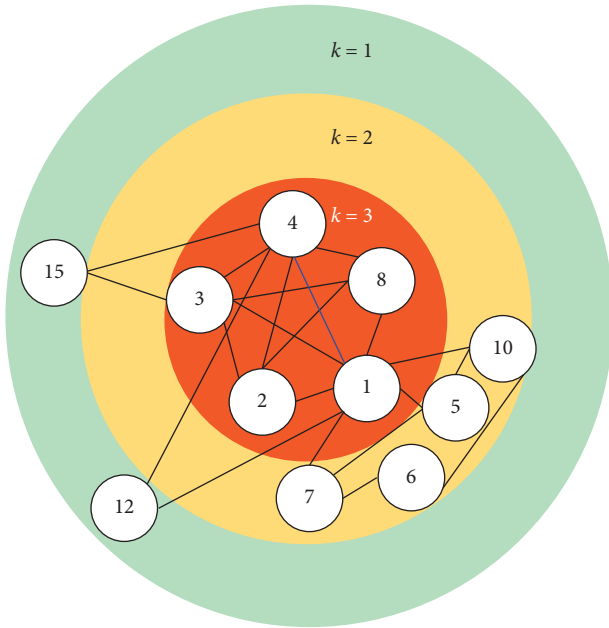

Figure 2: A k-shell layered diagram with nodes 9, 11, 13, and 14 deleted.

Table 1: Basic topological features of the four networks.

Networks	N	M	$<k>$	C
Lesmis	77	254	6.597	0.573
Polbooks	105	441	8.400	0.488
Metabolic	453	2040	9.007	0.646
Net-science	1461	2742	3.754	0.694

$$\mathrm{AUC} = \frac{0.5n_1 + n_2}{n}. \tag{1}$$

We use the classic link prediction algorithm RA (RA has the best overall performance over the other 10 algorithms [10]) and the new algorithm L3 (L3 performs best on the protein network [18]) in the experiment. For any two nodes x, y in the RA method, the possibility of link existence is calculated as follows:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}, \tag{2}$$

where $\Gamma(x)$ is the set of neighbor nodes of node x and $k(z)$ is the degree of node z.

For any two nodes x, y in the L3 method, the possibility of link existence is calculated as follows:

$$S_{xy} = \sum_{u,v} \frac{a_{xu} a_{uv} a_{vy}}{\sqrt{k(u)k(v)}}. \tag{3}$$

Among them, if there is a connection between node x and u, $a_{xu} = 1$; otherwise, it is 0. $k(u)$ is the degree of node u.

We divide the test set ratio in each network into 10%, 20%, 30%, 40%, and 50%. We take the largest connected subgraph of the test set and perform 20 independent experiments on each divided dataset. The obtained AUC are averaged. The AUC results of the two algorithms under this improved method are shown in Figure 3. RA* and L3* represent the results of using RA and L3 algorithms after applying the proposed method to delete some nodes.

Obviously, the AUC of L3* and RA* is higher than that of L3 and RA. Therefore, we deleted the nodes that are in the 1-shell and the nodes that are in 2-shell that are not linked to the x-shell ($x > 3$) in the above network, and subsequently using RA and L3 algorithms can improve the AUC values.

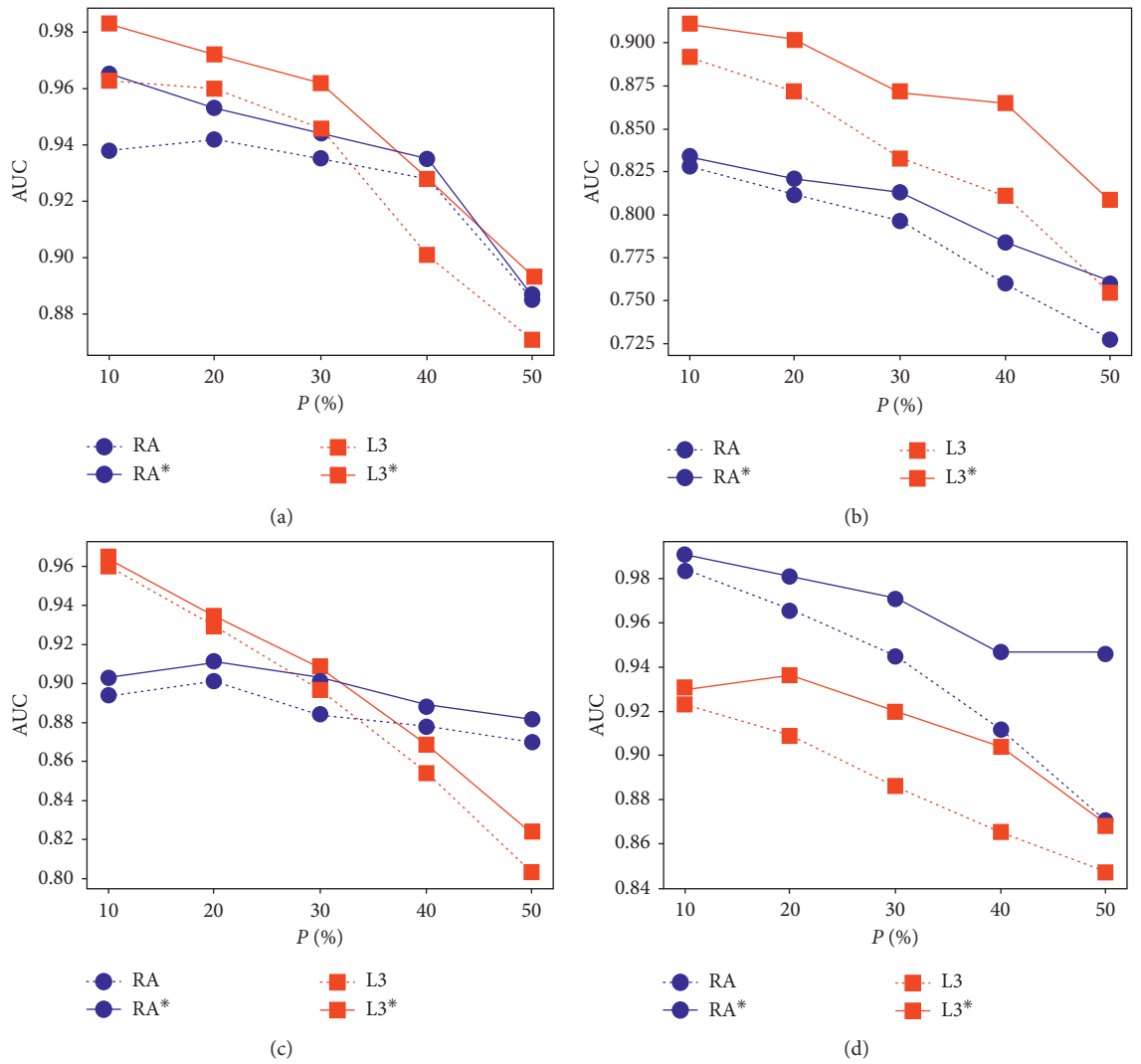

Figure 3: AUC result of four networks. The horizontal axis represents the proportion of the test set to the number of network edges, whereas the vertical axis denotes the AUC value. The blue dotted line represents the AUC of RA, and the blue solid line represents the AUC of RA*. The red dotted line represents the AUC of L3, and the red solid line represents the AUC of L3*. (a) Lesmis. (b) Polbooks. (c) Metabolic. (d) Net-science.

This finding indicates that the proposed method can improve the accuracy of link prediction.

The present study still requires further improvements, including the following: (1) determining whether all nodes that are in 1-shell should be deleted, (2) finding new methods for deleting nodes without affecting network connectivity to improve the accuracy of link prediction, and (3) determining whether the method for deleting nodes is suitable for other link prediction algorithms.

Our proposed method of deleting nodes provides an idea for dealing with large networks. However, we also found that some networks use k-shell layering with only one or a few layers. Deleting the outermost layer could possibly cause a large number of nodes to be deleted, which will affect the experimental results. Some other networks do not have 1-shell and 2-shell layers. These situations are important for us to solve.

4. Conclusion

This study proposes a method for improving the accuracy of link prediction. Before using the link prediction algorithm, the network is layered using k-shell, and the nodes in the 1-shell (outermost shell) and the 2-shell (secondary outer shell) that are not connected to those in the high-shell are deleted. The RA and L3 algorithms are used to conduct experiments on four real network datasets in different fields. The results show that the accuracy of the algorithms increases after deleting the abovementioned nodes, thereby verifying the effectiveness of the proposed method.

Data Availability

Data are available in the following links: http://www-personal.umich.edu/~mejn/netdata/, https://deim.urv.cat/~alexandre.arenas/data/welcome.htm, and http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been supported by the National Key Research and Development Program under No. 2018YFB2100100, National Natural Science Foundation of China under Grant No. 62066048, and Postdoctoral Science Foundation of China under No. 2020M673312.

References

[1] L. Getoor and C. P. Diehl, "Link mining," *Acm Sigkdd Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.

[2] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[3] B. Moradabadi and M. R. Meybodi, "Link prediction in fuzzy social networks using distributed learning automata," *Applied Intelligence*, vol. 47, no. 3, pp. 837–849, 2017.

[4] C. Dai, L. Chen, B. Li, and Y. Li, "Link prediction in multi-relational networks based on relational similarity," *Information Sciences*, vol. 394-395, pp. 198–216, 2017.

[5] H. Yu, P. Braun, M. A. Yildirim et al., "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, no. 5898, pp. 104–110, 2008.

[6] J. Kleinberg, "Analysis of large-scale social and information networks," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, 20120378 pages, 2013.

[7] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *The Journal of Mathematical Sociology*, vol. 1, no. 1, pp. 49–80, 1971.

[8] P. Jaccard, "Bulletin de la société vaudoise des sciences naturelles," *Etude Comparative de la Distribution florale dans une Portion des Alpes et des Jura*, vol. 37, pp. 547–579, 1901.

[9] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[10] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[11] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[12] L. Lü, C. H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, 46122 pages, 2009.

[13] S. Liu, X. Ji, C. Liu, and Y. Bai, "Extended resource allocation index for link prediction of complex network," *Physica A: Statistical Mechanics and Its Applications*, vol. 479, pp. 174–183, 2017.

[14] X. Li, S. Liu, H. Chen, and K. Wang, "A potential information capacity index for link prediction of complex networks based on the cannikin law," *Entropy*, vol. 21, no. 9, 863 pages, 2019.

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

[16] D. J. Klein and M. Randić, "Resistance distance," *Journal of Mathematical Chemistry*, vol. 12, no. 1, pp. 81–95, 1993.

[17] F. Fouss, A. Pirotte, J.-m. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 355–369, 2007.

[18] I. A. Kovács, K. Luck, K. Spirohn et al., "Network-based prediction of protein interactions," *Nature Communications*, vol. 10, no. 1, pp. 1–8, 2019.

[19] M. Kitsak, L. K. Gallos, S. Havlin et al., "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[20] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, ACM, New York, NY, USA, 1993.

[21] V. Krebs, "Books about US politics network dataset," 2004, http://www.orgnet.com/.

[22] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical Review E*, vol. 72, no. 2, 27104 pages, 2005.

[23] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, 36104 pages, 2006.

[24] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.