Advances in Engineering Research (AER), volume 130
5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)

Link Prediction via Extended Resource Allocation Index

LongjieLi[1, *], ShenshenBai[2], ShiyuYang[1], LongyuQu[1]and YiweiYang[1]

[1]ChinaInformationTechnologySecurityEvaluationCenter, Beijing 100085, China

[2]Lanzhou Vocational Technical College, Lanzhou 730070, China country

[*]corresponding author, email: li.longjie@hotmail.com

Keywords: complex network, link prediction, resource allocation, quasi-local index

Abstract. Link prediction is an important branch of complex network analysis, which can identify the missing or future links in a network. In this paper, a new link prediction method is presented, inspired by the ideas of both resource allocation index and quasi-local indices, to estimate the likelihood of existing a link between two unconnected nodes. To evaluate the prediction accuracy of the new index, we conduct experiments on five real-world networks compared with five famous indices. The results show that our new index outperforms the five baselines on the five networks.

Introduction

Researchers have found that many complex systems in real-world can be described as complex networks[1, 2], for example, social networks, biological networks, and commerce networks, in which nodes and links (or edges) represent individuals and their relationships, respectively. The research of complex networks has attracted many scholars and a sea of research achievements have been presented [3, 4, 5]. Among the various studies, *link prediction* is a very important topic and has received sustained attention [6, 7, 8].

Link prediction is a fundamental task in complex network analysis and has a wide range of applications in recommend system, information retrieval, and bioinformatics, etc. The purpose of link prediction is to find or predict the links which are missed or will appear in a network. In real-world, the available networks are usually incomplete [9, 10], for instance, protein-protein interaction networks. Therefore, link prediction has critical values since it can find the missing links for those networks. Moreover, link prediction can help us to understand the evolution process of networks [11, 12, 13].

So far, many link prediction methods have been proposed by researchers from different disciplines [6, 8]. Among them, a series of methods are designed based on similarities between nodes, which are the so-called *similarity-based methods*. Those methods assume that a link is more likely existent between two unconnected nodes if they have higher similarity [6]. Thus, the key problem is to define a sound similarity index between nodes. In general, only the structure information of a network can be obtained. Hence, similarity indices based on structure information are the concern of researchers [14, 15]. One group of similarity indices is based on common neighbors. The well-known *Common Neighbors* (CN) index simply sums the shared neighbors of a pair of nodes [7]. The *Jaccard* index [7] and *Salton* index [16] are two normalizations of CN index, which take the degrees of endpoints into account. Besides, the *Adamic-Adar* (AA) index [17] and *Resource Allocation* (RA) index [18] improve the CN index by penalizing high degree neighbors. Several works have proved that the RA index achieves the best results among the aforementioned indices [6, 19]. The common neighbor-based indices merely use the local structure of a network. Therefore, their performance in efficiency is very high, but performance in prediction is relatively poor. On the other hand, there are some other similarity-based methods, *Kate* index [20], *SimRank* [21] index and *Average Commute Time* (ACT) index [22], to name a few, which compute similarities between nodes based on the global structure of a network. Consequently, these methods suffer from high computational complexity. To balance the prediction accuracy and computational complexity between the two classes indices, the third class of similarity indices has been studied. This class is based on the quasi-local structure of a network, including *Local Path* (LP) index [23], *FriendLink* index [24] and *Local Random Walk* (LRW) index [25], etc.

In this paper, an extended resource allocation index (ERA for short), which uses the ideas of both RA index and LP index for reference, is proposed. Given two unconnected nodes, namely *u* and *v*. In RA index, node *u* can only send its resource via their common neighbors to node *v*. If *u* and *v* have no common neighbors, their similarity score equals to zero. However, the ERA index employs the paths with both length 2 and 3 between two unconnected nodes to calculate their similarity. In ERA index, node *u* can send its resource to node *v* through paths with length 2 and 3; intermediate nodes of those paths are transmitters. We compared the proposed similarity index with five baselines on five networks. Experimental results show that the ERA index outperforms the compared methods.

Problem and Metric

Consider an undirected and unweighted network $G(V,E)$, where$V$ and E denote the node set and link set, respectively. In this paper, multi-links and self-loops are not allowed in G. Let Ube the universal set, which contains all$(|V|*(|V|-1))/2$ possible links, where |V| is the number of nodes. The set of nonexistent links is $U\backslash E$. Suppose there are some missing links or future links in $U\backslash E$. The task of link prediction is to identify these links based on the observed network information. To solve this problem, one similarity-based method assigns a similarity score to each nonexistent link, and then sorts all nonexistent links according to their scores in descending order. The links at the top are considered as the missing or future links.

To estimate the accuracy of prediction methods, we randomly partition the link set E into two parts. The first part is the training set, denoted by E^T; while the second part is the probe set, denoted by E^P. Obviously, $E=E^T\cup E^P$ and $E^T\cap E^P=\emptyset$. To decrease the random bias, in this paper, we conducted 100 independent experiments for each individual network. In each run, 10% links are randomly extracted to build the probe set, while the remaining 90% links are used as training set.

The standard metric, *AUC*[26], is adopted to quantify the accuracy of prediction methods. The AUC value is the probability that the similarity score of a randomly selected links from probe set (E^P) is higher than that of a randomly selected links from nonexistent link set (U\E). In the implementation, we preforrm *n* independent comparisons. Let $n_1$denote the times that the link in E^Phas a higher score, and n_2 be the times that the link in E^P has the same score with the link in U\E. Then, the AUC value is defined as $AUC=(n_1+n_2)/n$. In our experiments, *n* is set to be 10,000.

Baselines and Datasets

In this paper, five famous similarity indices are used as baselines for the purpose of performance comparison, and their definitions are given as follows.

(1) Common Neighbors (CN) index

$$CN(u,v)=|N(u)\cap N(v)| \tag{1}$$

where N(u)is the neighbor set of node *u*.

(2) Adamic-Adar (AA) index

$$AA(u,v)=\sum_{w\in N(u)\cap N(v)}\frac{1}{\log(k_w)} \tag{2}$$

where$k_w$denotes the degree of node *w*.

(3) Resource Allocation (RA) index

$$RA(u,v)=\sum_{w\in N(u)\cap N(v)}\frac{1}{k_w} \tag{3}$$

(4) Jaccard (JA)index

$$JA(u,v)=\frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|} \tag{4}$$

(5) Local Path (LP) index

$$LP(u,v)=|P^2(u,v)|+\epsilon|P^3(u,v)| \tag{5}$$

where $P^2(u,v)$ and $P^3(u,v)$are the path sets between nodes u and v with length 2 and 3 respectively. ϵ is a free parameter to tune the influence of paths with length 3. In Ref. [23], the authors suggested to give a small positive value to ϵ. Thus, in our experiments, we set $\epsilon = 0.001$.

Five real-world networks are used as test datasets in this paper. They are: CE (neural network of C.elegans) [3], Karate (social network of a karate club) [27], NS (collaboration network between network scientists) [28], USAir (US airline network) [29], and Email (a network of email interchanges) [30]. All datasets are treated as undirected and unweighted networks in this paper. The basic topological features of these networks are listed in Table 1. Since some networks are not connected, we use the giant components of those networks.

Table 1. The basic topological features of the giant components of five datasets. $|V|$: node number; $|E|$: edge number; $\overline{k}$: average degree; $\overline{d}$: average shortest distance; C: clustering coefficient [3]; r: assortative coefficient [31]; H: degree heterogeneity, $H = \overline{k^2}/\overline{k}$.

Networks	|V|	|E|	$\overline{k}$	$\overline{d}$	C	r	H
CE	297	2148	14.465	2.455	0.292	-0.163	1.801
Karate	34	78	4.588	2.408	0.571	-0.476	1.693
NS	379	914	4.823	6.042	0.741	-0.082	1.663
USAir	332	2126	12.807	2.738	0.625	-0.208	3.464
Email	1133	5451	9.622	3.606	0.220	0.078	1.942

The New Method

As mentioned above, RA is an excellent common neighbor-based index. However, only using local structure information makes it losing the influence of other structure information. In particular, for two unconnected nodes, if they do not share any neighbor, RA will assign zero similarity score to them. Nevertheless, there may exist a non-observed link between these two nodes. To overcome this weakness of RA index, an extended RA index, *ERA* index, is proposed in this paper. Considering the advantage of quasi-local methods, the ERA index uses quasi-local structure information to implement the idea of RA.

In the original RA index, resource of node u is send to node v through paths between them with length 2; the common neighbors playing the role of transmitters. In ERA index, resource of node u can be send to node v through paths with length 2 as well as 3; all intermediate nodes in those paths play the role of transmitters. The definition of ERA is

$$ERA(u,v) = \sum_{i=2}^{3} \sum_{p \in P^i(u,v)} \prod_{M(p)} \frac{1}{k_w} \tag{6}$$

where $\mathrm{P}^{\mathrm{i}}(\mathrm{u},\mathrm{v})$ denotes the set of paths between u and v with length i, and $M(p)$ is the set of intermediate nodes of pathp. In addition, we can define ERA in another form, it is

$$ERA(u,v) = RA(u,v) + \sum_{p \in P^3(u,v)} \prod_{M(p)} \frac{1}{k_w} \tag{7}$$

In the following, we show the computation of RA and ERA by taking the toy network in Fig. 1 as an example. The respective values of $RA(a,b)$ and $ERA(a,b)$ are calculated as follows:

$$RRA(a,b) = \frac{1}{k_b} + \frac{1}{k_f} = \frac{1}{2} + \frac{1}{3} = \frac{5}{6},$$

$$ERA(a,b) = RA(a,b) + \frac{1}{k_d} \times \frac{1}{k_e} + \frac{1}{k_f} \times \frac{1}{k_g} = \frac{5}{6} + \frac{1}{4} + \frac{1}{6} = \frac{5}{4}.$$

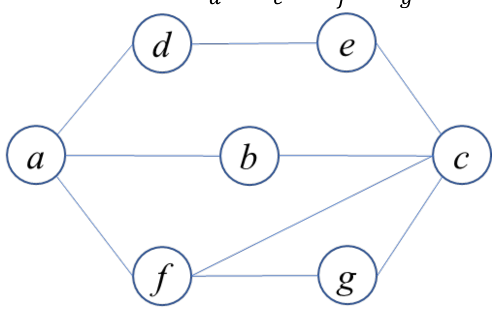

Fig. 1. An example network

Experimental Results

In this section, we perform experiments on five real-world networks, compared with five baselines in terms of accuracy measured by AUC. Table 2 shows the results. In Table 2, each AUC value is the average of 100 independent runs, numbers in brackets are standard deviations, and the best result for each network is highlighted in boldface.

Clearly, the ERA index always achieves the best prediction performance on all five networks, especially on CE, Karate, and Email. Compared with RA, the AUC values are improved from 0.8628 to 0.8986 on CE, from 0.7324 to 0.8009 on Karate, and from 0.8425 to 0.9039 on Email. On both NS and USAir, although the improvements are not so sharp, ERA still performs better than RA. Since the ERA index uses more structure information than RA index when computing the similarity score of a pair of nodes, it outperforms RA in terms of accuracy. In addition, ERA is superior to LP as shown in Table 2, although both ERA and LP are quasi-local index. The reason is ERA adopts the idea of resource propagation, which is more effective than simply counting the number of paths in LP.

Table 2. Accuracy of each link prediction method on five networks in terms of AUC.

Methods	CE	Karate	NS	USAir	Email
CN	0.8416(0.0155)	0.6874(0.0949)	0.9513(0.0178)	0.9353(0.0090)	0.8408(0.0086)
AA	0.8582(0.0145)	0.7240(0.1011)	0.9546(0.0181)	0.9467(0.0083)	0.8429(0.0088)
RA	0.8628(0.0144)	0.7324(0.1031)	0.9548(0.0181)	0.9526(0.0079)	0.8425(0.0089)
JA	0.7943(0.0119)	0.6151(0.0868)	0.9430(0.0168)	0.8966(0.0111)	0.8408(0.0100)
LP	0.8588(0.0134)	0.7158(0.0820)	0.9513(0.0195)	0.9281(0.0120)	0.8978(0.0102)
ERA	0.8986(0.0108)	0.8009(0.0895)	0.9567(0.0197)	0.9528(0.0101)	0.9039(0.0093)

In summary, the ERA index is an efficient quasi-local link prediction method, which achieves the best prediction accuracy compared with four common neighbor-based indices (i.e., CN, AA, RA, Jaccard) and one quasi-local index (i.e., LP).

Conclusion

In this paper, a new quasi-local index is proposed to perform the task of link prediction. This new index implements resource allocation via paths connecting two nodes with length 2 and 3; the intermediate nodes in those paths play the role of transmitters. From the experimental results, we can clearly observe that our new index achieves more accurate prediction than baselines (i.e., CN, AA, RA, Jaccard, and LP). The new index only uses quasi-local structure information of a network, so its computational complexity equals to LP. Therefore, it is feasible on large network.

References

[1] Katy Börner, Soma Sanyal, Alessandro Vespignani. Network science[J]. Annual review of information science and technology, 2007 41(1):537–607.

[2] Sheng-Jun Wang, Zhen Wang, Tao Jin, Stefano Boccaletti. Emergence of disassortative mixing from pruning nodes in growing scale-free networks[J]. Scientific reports, 2014 4:7536.

[3] D. J. Watts, S. H. Strogatz. Collective dynamics of small-world networks[J]. Nature, 1998 393(6684):440–442.

[4] Réka Albert, Albert-László Barabási. Statistical mechanics of complex networks[J]. Reviews of Modern Physics, 2002 74:47–97.

[5] M. E. J. Newman. The structure and function of complex networks[J]. SIAM review, 2003 45(2):167–256.

[6] LinyuanLü, Tao Zhou. Link prediction in complex networks: A survey[J]. Physica A, 2011 390(6):11501170.

[7] David Liben-Nowell, Jon Kleinberg. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007 58(7):1019–1031.

[8] Peng Wang, Baowen Xu, et al. Link prediction in social networks: the state-of-the-art[J]. Science China Information Sciences, 2015 58(1):1–38.

[9] Michael P. H. Stumpf, Thomas Thorne, et al. Estimating the size of the human interactome[J]. Proceedings of the National Academy of Sciences, 2008 105(19):6959–6964.

[10] Dongbo Bu, Yi Zhao, et al. Topological structure analysis of the protein-protein interaction network in budding yeast[J]. Nucleic Acids Research, 2003 31(9):2443.

[11] Wen-Qiang Wang, Qian-Ming Zhang, Tao Zhou. Evaluating network models: A likelihood analysis[J]. EPL (Europhysics Letters), 2012 98(2):28004.

[12] Qian-Ming Zhang, Xiao-Ke Xu, et al. Measuring multiple evolution mechanisms of complex networks[J]. Scientific Reports, 2016 5:1035.

[13] Carlo V.Cannistraci, Gergo Alanis-Lobato, Timothy Ravasi. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks[J]. Scientific Reports, 2015 3(4):1613.

[14] Guo-Dong Lyu, Chang-Jun Fan, et al. Predicting missing links via structural similarity[J]. International Journal of Modern Physics B, 2015 29(15):1550095.

[15] Jinxuan Yang, Xiao-Dong Zhang. Predicting missing links in complex networks based on common neighbors and distance[J]. Scientific Reports, 2016 6:3820.

[16] Gerard Salton, Michael J. McGill. Introduction to modern information retrieval[M]. McGraw-Hill, New York, 1983.

[17] Lada Adamic, Eytan Adar. Friends and neighbors on the web[J]. Social Networks, 2003 25(3):211–230.

[18] Tao Zhou, LinyuanLü, Yi-Cheng Zhang. Predicting missing links via local information[J]. The European Physical Journal B, 2009 71(4):623–630.

[19] Peng Zhang, Xiang Wang, et al. Measuring the robustness of link prediction algorithms under noisy environment[J]. Scientific reports, 2016 6:18881.

[20] Leo Katz. A new status index derived from sociometric analysis[J]. Psychometrika, 18(1):39–43, 1953.

[21] Glen Jeh, Jennifer Widom. Simrank: a measure of structural-context similarity[C]. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002 538–543.

[22] Francois Fouss, Alain Pirotte, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Transactions on knowledge and data engineering, 2007 19(3).

[23] LinyuanLü, Ci-Hang Jin, Tao Zhou. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E, 2009 80:046122.

[24] Alexis Papadimitriou, Panagiotis Symeonidis, YannisManolopoulos. Fast and accurate link prediction in social networking systems[J]. Journal of Systems and Software, 2012 85(9):2119–2132.

[25] Weiping Liu, LinyuanLü. Link prediction based on local random walk[J]. EPL (EurophysicsLetters), 2010 89(5):58007.

[26] J. A. Hanley, B. J. Mcneil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases.[J]. Radiology, 1983 148(3):839–843.

[27] Wayne W. Zachary. An information flow model for conflict and fission in small groups[J].

Journal of Anthropological Research, 1977 33(4):473.

[28] Mark E. J. Newman. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006 74(3):036104.

[29] Vladimir Batagelj, Andrej Mrvar. Pajek datasets, 2006.

[30] R. Guimerà, L. Danon, et al. Self-similar community structure in a network of human interactions[J]. Physical Review E, 2003 68:065103.

[31] Mark. E. J. Newman. Mixing patterns in networks[J]. Physical review E, 2003 67:026126.