Physica A 492 (2018) 1523–1530

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Link prediction via significant influence

Yujie Yang [a,b], Jianhua Zhang [a,*], Xuzhen Zhu [a,*], Lei Tian [a]

[a] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Mailbox NO.92, Beijing, 100876, PR China

[b] Institute of Computer and Information Technology, Henan Normal University, Xinxiang 453007, PR China

HIGHLIGHTS

- A novel link prediction index significant influence (SI) is proposed.
- The strong and weak influences in transferring resources are defined.
- SI models significant influence by distinguishing the strong influence from the weak.
- The proposed index can achieve the better performance than the traditional indices.

ARTICLE INFO

Article history:
Received 19 April 2017
Received in revised form 29 September 2017
Available online 16 November 2017

Keywords:
Link prediction
Significant influence
Similarity
Complex network

ABSTRACT

In traditional link prediction, many researches assume that endpoint influence, represented by endpoint degree, prefers to facilitate the connection between big-degree endpoints. However, after investigating the network structure, it is observed that influence is determined by the relations built through the paths between endpoints instead of the endpoint degree. Strong relations connecting the other endpoint through short paths, especially through common neighbors, can bring in more powerful influence, and in contrast, those relations through long paths obviously generate weak influence. In this paper, a novel link prediction index SI is proposed, which deliberately models the significant influence by distinguishing the strong influence from the weak. After comparison with main stream baselines on 12 benchmark datasets, the results suggest SI effectively improve the link prediction accuracy.

1. Introduction

Recently, many studies investigate topological features and reveal network functions for comprehensive understanding the essential characters of complex networks [1–5]. Link prediction is put forward to solve the related problems in some researches and has attracted more attentions [6,7]. Link prediction indicates how to utilize the information of endpoint and network structure to predict the connecting possibility of two unconnected endpoints. It can be applied in many fields such as exploring protein-to-protein interactions [8,9], studying the potential mechanism which drives co-authorship evolution [10], reconstructing airline networks [11], recommending friends [12,13] and promoting e-commerce scales [14,15], etc.

The similarity-based method of link prediction is defined based on the network structure and is more applicable than other methods. The similarity index is usually modeled to describe the probability of finding the missing and future

* Corresponding authors.
E-mail addresses: jhzhang@bupt.edu.cn (J. Zhang), zhuxuzhen@bupt.edu.cn (X. Zhu).

https://doi.org/10.1016/j.physa.2017.11.078
0378-4371/

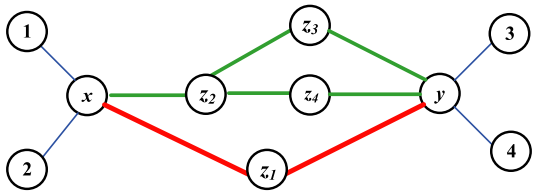

Fig. 1. Illustration of the relationship between the initial endpoint and the target endpoint, where the red line indicates the two-hop path, green line is the three-hop path and the blue one expresses the path disable to connect the other endpoint. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

links [7,16]. It is a problem that huge expenses are caused by hardly extracting the attributes from endpoints in link prediction [17,18]. Considering the topological similarity based on the network structure, mainstream methods for link prediction can be divided into three classes. The first class is called global index, such as Katz Index [19], which uses the global structural information to calculate the topological similarity of the endpoints. Unfortunately, global index suffers high computational complexity. The second class is proposed on local structure of network. Traditional local indices model the similarity by counting the number of common neighbors (CN) [20], or setting penalizing parameter to punish the large-degree endpoints, such as Salton Index [21], Sorensen Index [22], Hub Promoted Index [23], Leicht–Holme–Newman Index [24] and so on. Adamic–Adar Index (AA) [25] and Resource Allocation Index (RA) [26] penalize the large-degree common neighbors on the basis of CN index. Compared with the global index, the local indices have the lower complexity but suffer the poor performance. The third class focuses on quasi-local structures of network in order to get the compromise between performance and complexity. The Local Path Index (LP) [26,27] considers the two and three hops paths but ignores the longer paths. The Local Random Walk (LRW) and the Superposed Random Walk (SRW) are the similarity indices based on random walk [28]. LRW just considers the process of limited number of steps, while SRW gives the nodes nearby more opportunities to be connected to the target node. Some recent works propose new methods based on these traditional indices for improving performance. Zhu et al. [29] supposes that paths consisting of small-degree nodes contribute more in the similarity between endpoints and propose a significant path index by using the intermediate node-degree to calculate the similarity. Liu et al. [30] filters out the redundant links in the network to improve the accuracy of the k-shell method from the perspective of spreading dynamics. Zeng [31] presents an index of common neighbor plus preferential attachment to estimate the possibility of the link existence. Ahmed et al. [32] presents a fast algorithm via random walks in temporal networks.

Previous researches assume that endpoint influence helps unconnected endpoints to connect each other in the future. They, however, simply regard the endpoint degree as the effective influence, based on which one endpoint attracts another unconnected endpoint in the future. Based on the Three Degree of Influence Rule [33], we find the influence of endpoint is eventually determined by the paths from it to its target endpoint, but the endpoint degree. For example, although possessing many relationships in the social network, two strangers are more likely to know each other through a common friend but the indirect chain of friends, i.e., more co-friends mean more effective connections, promoting two people to know each other and to be more similar. Moreover, links constructing the endpoint degree possess different abilities in transferring influence between two endpoints, namely, some links deliver more by common neighbors, some deliver less by long paths with three or more hops and the others even cannot connect the target endpoint anyway. Accordingly, for an endpoint, the ability of a short path contributing more in future connection should be called strong relation, and the ability of a long path should be called weak relation oppositely. Obviously, significant influence holds more strong relations and less weak ones.

A simple example to illustrate the strong and weak relations in the network is shown in Fig. 1. There are three different paths marked with disparate colors between initial endpoint x and target endpoint y. In the red two-hop path $x - z_1 - y$, endpoint x connects directly with y through a common neighbor z_1. This path can produce the strong relation between x and y because of the short length. The green three-hop paths, both built by two intermediate nodes between x and y, are regarded as the weak relations of the influence, which is smaller than the short path's. Moreover, some blue paths where x are disconnected to y contribute the least influence. Fig. 1 at the same time exemplifies the fact that the influence is determined by the relations represented by paths instead of degree by links, i.e., the influence delivered by the two paths $x - z_2 - z_3 - y$ and $x - z_2 - z_4 - y$ cannot be simply delivered by the one single link $x - z_2$. So the endpoint degree is inappropriate to be modeled as effective influence. Obviously, we believe more strong relations and less weak ones constitute the significant influence, which promotes the future connection and similarity of the endpoints.

In this paper, through emphasis of the strong relations and penalization of the weak, we propose a novel link prediction index via modeling the significant influence (SI). In comparison experiments with main stream indices on 12 benchmark datasets, the results exhibit the excellent improvement in the link prediction accuracy. The remainder of this paper is organized as follows. Section 2 defines the SI index for link prediction and some baselines for comparison in complex network. The datasets and metrics for experiment are given in Section 3. We discuss the results in Section 4 and conclude the whole paper in Section 5.

2. Method

2.1. Definition

Above all, the definitions of strong and weak relations and significant influence are given as blow.

Definition 1. In an undirected and unweighted network $G(V, E)$, the relation of endpoints is built through the paths between them. Between endpoints x and y, the short path, especially two-hop path, is represented as strong relation, whereas weak relation happens when x connects with y through long path. When endpoint x has more strong relations and less weak ones with y, we can think there exists significant influence between them. According to Three Degrees of Influence Rule, the strong relation generates strong influence, which can be defined as,

$$I_{xy}^{strong}(t) = |\Gamma(x) \cap \Gamma(y)| * [\sum_{\tau=1}^{t} \pi_{xy}(\tau) + \sum_{\tau=1}^{t} \pi_{yx}(\tau)] \tag{1}$$

where $\pi_{xy}(\tau)$ denotes the transfer probability from x to y and $|\Gamma(x) \cap \Gamma(y)|$ indicates the number of common neighbors between endpoints x and y. The influence contributed by the weak relation is called weak influence, which can be defined as,

$$I_{xy}^{weak}(t) = \sum_{l=3}^{\infty} \left|paths_{x,y}^{\langle l \rangle}\right| * [\sum_{\tau=1}^{t} \pi_{xy}(\tau) + \sum_{\tau=1}^{t} \pi_{yx}(\tau)] \tag{2}$$

where $|paths_{x,y}^{\langle l \rangle}|$ denotes the number of paths between x and y with length l. We believe that paths with three or more hops bring in the weak relation, so the path length is set from $l = 3$ to the infinity in the Eq. (2).

Apparently, more strong influences and less weak ones are preferred when transferring resources, so a penalizing parameter α is adopted to punish the weak influence from the long paths. Furthermore, the influences between x and y are directly presented as the similarity, and the significant influence can be defined as below.

Definition 2. In an undirected and unweighted network $G(V, E)$, the significant influence possessed by endpoints x and y should emphasize the strong influence and penalize the weak influence between them as,

$$\begin{aligned} s_{xy}^{SI}(t) = {} & |\Gamma(x) \cap \Gamma(y)| * [\sum_{\tau=1}^{t} \pi_{xy}(\tau) + \sum_{\tau=1}^{t} \pi_{yx}(\tau)] \\ & + (\sum_{l=3}^{\infty} |paths_{x,y}^{\langle l \rangle}|)^{\alpha} * [\sum_{\tau=1}^{t} \pi_{xy}(\tau) + \sum_{\tau=1}^{t} \pi_{yx}(\tau)]. \end{aligned} \tag{3}$$

Definition 2 defines SI index which highlights the importance of the strong influence, mainly from common neighbors, by applying a punishment parameter α to restrict the weak influence from three or more hops paths between the initial and the target endpoints. α plays the penalization role in SI, so it should vary in $(-\infty, +1]$.

2.2. baselines

For comparison, five classical indices are introduced as blow:

(1) Common Neighbors (CN) [20]. CN is the simplest similarity based index which believes the more common neighbors between the two endpoints the more similar they are. The similarity of two endpoints is defined by the number of common neighbors,

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)| \tag{4}$$

where $\Gamma(x)$ indicates the neighbor nodes set of endpoint x and $\Gamma(x) \cap \Gamma(y)$ denotes the set of common neighbors of endpoints x and y.

(2) Adamic–Adar (AA) [25]. AA gives a punishment weight to the large-degree of common neighbor based on CN. The weight equals to the reciprocal of logarithm of the large-degree endpoint. It is defined as,

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(k_z)} \tag{5}$$

where k_z denotes the degree of common neighbor z.

(3) Resource-Allocation (RA) [26]. RA just considers the reciprocal of common neighbor's degree, penalizing the large-degree of common neighbors, similar as AA index. It is defined as,

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{6}$$

(4) Local Path (LP) [26,27]. LP is the similarity index based on local path, considering the number of the two and three hops paths. For restraining the influence of three-hop path, LP sets an adjustable parameter ε as,

$$s_{xy}^{LP} = A^2 + \varepsilon \cdot A^3 \tag{7}$$

where A denotes the adjacent matrix of the network and ε represents the penalization parameter from 0 to 1.

(5) Superposed Random Walk (SRW) [28]. SRW not only considers the local paths but also sets the distribution of node-degree as the initial resource. Superposing the t steps random walks as,

$$s_{xy}^{SRW}(t) = q_x \sum_{\tau=1}^{t} \pi_{xy}(\tau) + q_y \sum_{\tau=1}^{t} \pi_{yx}(\tau) \tag{8}$$

where $q_x = \frac{k_x}{2|E|}$ denotes the initial resource distribution. $\pi_{xy}(\tau)$ expresses the transfer probability starting from endpoint x to y through t walks. We adopt $\pi_x(t+1) = P^T \pi_x(t)$, $t \geq 0$ and $\pi_x(0) = e_x$. e_x represents an $N \times 1$ vector with the xth element as 1 and the others as 0. The transition probability matrix is P, with $P_{xy} = \frac{a_{xy}}{k_x}$, where k_x denotes the endpoint degree. a_{xy} equals to 1 if endpoints x and y are connected, 0 otherwise. T represents the matrix transposition.

3. Experiments

In an unweighted and undirected network $G(V, E)$, V and E represent the set of nodes and links respectively. The links set E is randomly divided into two parts: the training set E^T treated as known information and the testing set E^P used for prediction. The division should guarantee the connectivity in E^T. Clearly, $E^T \cup E^P = E$ and $E^T \cap E^P = \phi$. We suppose the universal set as U which contains all the $\frac{|V| \times (|V|-1)}{2}$ links. Then the nonexistent links set can be represented as $U \setminus E$. The purpose of link prediction is to predict the probabilities of links in E^P and $U \setminus E$. In our experiment, the training set E^T contains 80% links and the rest 20% links are in E^P. Before showing the results of experiments, the datasets and the metrics are introduced in advance.

3.1. Data

We use 12 real networks as the experiment datasets.[1] Before the experiment, the directed graph is firstly converted into the undirected and then the loops and multi-links are deleted.

The datasets include: (1) US Air97 (USAir) [34]. USAir is the aviation network of USA. (2) Yeast PPI (Yeast) [35]. Yeast is the protein–protein interaction network which regards protein and their interaction as node and edge respectively. (3) Food Web of Florida ecosystem (FW) [36]. FW is the network of food chain in the rainy season of Florida. The predation relationship is described by the carbon exchange relation in this network. (4) Power [37]. Power is the American West electrical network. The node expresses generator, transformer and substation, while the edge represents the high voltage transmission line between nodes. (5) NetScience (NS) [38]. NS is the scientist cooperation network in which the node represents scientist and the edge is the cooperation relationship between scientists. (6) C.elegans (CE) [37]. CE is a neural network of the worm Caenorhabditis elegans. (7) E-mail network (Email) [39]. Email indicates the communication network which uses the email in University Rovira i Virgili (URV) in Tarragona, Spain. (8) Jazz [40]. Jazz denotes the network including the collaboration between jazz musicians. (9) Political blogs (PB) [41]. PB is the network constructed by the web pages of US political blogs. The edge represents the hyperlink of the web page. (10) Eurosis web mapping study (ES) [42]. ES describes a mapping network between scientists and social activists in 12 European countries. (11) Infectious (Infec) [43]. Infec is an infection network. (12) Slavko [44]. Slavko is the network involving the Facebook friendship of Slavko Žitnik.

The basic topological features of these networks are shown in Table 1 where $|V|$ and $|E|$ represent the numbers of nodes and links respectively. $\langle k \rangle$ is the average degree, $\langle d \rangle$ denotes the shortest average distance, C indicates the clustering coefficient [37], r expresses the assortativity coefficient [45] and H is the degree heterogeneity, defined as $H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$.

[1] Some data sets are freely downloadable from the following academic web sites: http://vlado.fmf.uni-lj.si/pub/networks/data, http://wiki.gephi.org/index.php?title=Datasets, http://lovro.lpt.fri.uni-lj.si/support.jsp, and http://www.linkprediction.org/index.php/link/resource/data.

Table 1
The basic topology features of twelve benchmark networks.

Network	$\|V\|$	$\|E\|$	$\langle d\rangle$	$\langle k\rangle$	C	r	H
USAir	332	2 126	2.73	12.81	0.749	−0.208	3.36
Yeast	2364	10 904	5.16	9.20	0.378	0.469	3.35
FW	128	2 075	1.77	32.42	0.334	−0.112	1.25
Power	4941	6 594	15.87	2.669	0.107	0.003	1.45
NS	1461	2 742	5.82	3.451	0.878	0.461	1.85
CE	453	2 025	2.66	8.94	0.655	−0.225	4.23
Email	1133	5 451	3.61	9.62	0.254	0.078	1.94
Jazz	198	2 742	2.24	27.7	0.633	0.02	1.4
PB	1222	16 717	2.51	27.36	0.360	−0.221	2.97
ES	1272	6 454	3.86	10.15	0.382	−0.012	2.46
Infec	410	2 765	3.63	13.49	0.467	0.226	1.39
Slavko	334	2 218	3.05	13.28	0.488	0.247	1.62

Table 2
AUCs on eleven datasets, averaged over ten divisions of datasets with $n = 100\,000$.

AUC	CN	AA	RA	LP	SRW	SI
USAir	0.938	0.950	0.956	0.931	0.954	**0.958**
Yeast	0.723	0.724	0.723	0.736	0.737	**0.737**
FW	0.613	0.615	0.620	0.629	0.716	**0.740**
Power	0.591	0.591	0.591	0.642	0.643	**0.643**
NS	0.940	0.940	0.940	0.940	0.944	**0.944**
CE	0.914	0.948	0.954	0.914	0.956	**0.961**
Email	0.844	0.846	0.846	0.893	0.907	**0.908**
Jazz	0.954	0.961	0.970	0.954	0.965	**0.971**
PB	0.915	0.918	0.919	0.930	0.939	**0.942**
ES	0.910	0.912	0.912	0.936	0.946	**0.946**
Infec	0.939	0.943	0.944	0.954	0.965	**0.966**
Slavko	0.941	0.945	0.946	0.944	0.951	**0.953**

3.2. Metric

A standard metric, AUC (Area Under the Receiver Operating Characteristic Curve) [46], is used to evaluate the prediction performance of our index. AUC is the probability that the score of a link randomly chosen in the testing set is higher than that of a nonexistent link. That is, we each time from the testing set E^P randomly select a link, and then from the nonexistent link set, $U \setminus E$, randomly select another link. If the score of link in E^P is higher than that of $U \setminus E$, then we get 1. If two scores are equal, then we get 0.5. After n times independent comparisons, if there are n' times that the links in testing set have higher scores and n'' times they are the same, the AUC can be presented as:

$$AUC = \frac{n' + 0.5n''}{n}. \tag{9}$$

Obviously, if all the scores are generated randomly, the value of AUC approximately equals to 0.5. The extent of AUC greater than 0.5 measures the extent to which the index is more accurate than the randomly selected method.

4. Results and discussions

The AUC curves of SI are illustrated in Fig. 2 under twelve independent datasets. The α axis represents the degree of punishment and the vertical axis indicates the accuracy of the SI index. In order to show the impact of the penalty factor on the accuracy of the index and better observe the changing trend of AUC, we set α to vary in a wider range of $[-2, 2]$. AUC varies continuously in all datasets, as shown in Fig. 2. The penalization parameter plays important role in prediction when $\alpha < 1$. Following this thought, the extent to which α is smaller than 1 indicates how seriously the weak influence is suppressed. At the same time, the suppression of the weak influence enhances the significant influence. For every dataset in Fig. 2, the optimal AUC is acquired when the absolute value of α is in $[0, 1]$, which verifies the intuition we have claimed that α does serve as penalization. In most of datasets, the curves all have a decrease trend after the optimal value of AUC, meaning that penalizing weak influence can improve the accuracy of the index. It is just because SI purposely emphasizes the significant influence of two-hop paths involving common neighbors between endpoints, and effectively punishes the weak influence of long paths in transferring resource. Accordingly it obtains the better performance than other indices.

To demonstrate the prediction ability, we give the performance of SI index with the optimal α value on twelve data sets, respectively. The average AUC values of SI and other five classical indices are shown in Table 2. We take $\varepsilon = 0.001$ in LP index and $t = 3$ in SRW index. As shown in Table 2, the performance of CN is the worst in all of the indices because it only considers the number of two-hop paths between two endpoints but ignores the influence of long paths. On the basis of CN,

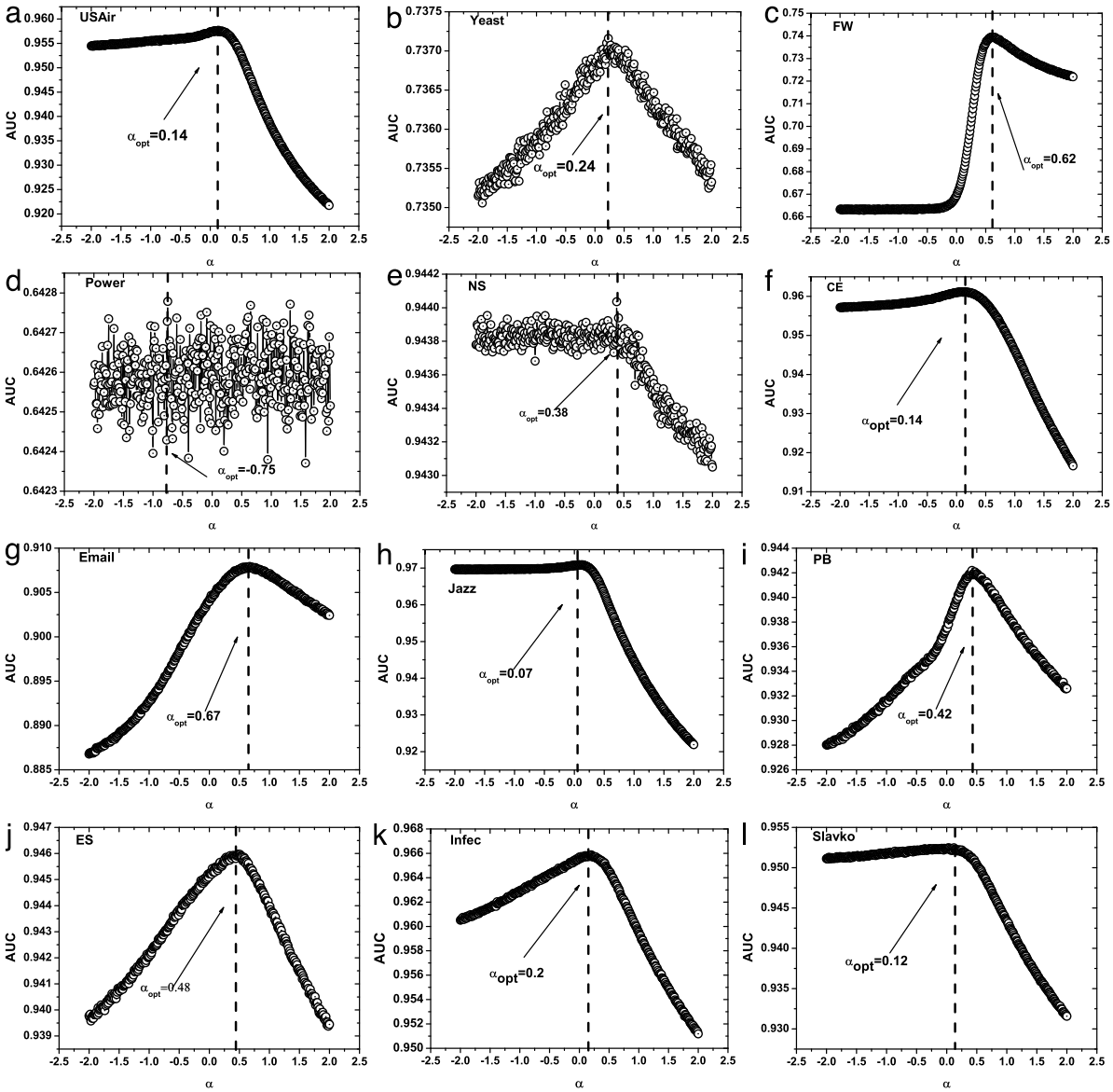

Fig. 2. The prediction performance of SI index on 12 benchmark networks with different values of α. For each network, the optimal values of α is presented inside the corresponding plot. Notice that, for every case, the optimal AUC is achieved when α is smaller than 1, indicating that to emphasize the significant influence and penalize the weak influence are effective in link prediction.

AA and RA achieve the improved performance, but still do not involve the influence of long paths. LP and SRW consider paths with two and three hops, and obtain the better performance than the indices just concerning two-hop paths. But they neither distinguish the different influences between the two-hop and multi-hop paths nor effectively restrict the weak influence in transferring resources. SI emphasizes the significant influence through the effective two-hop paths by penalizing the weak influence of three or more hops paths and achieves the optimal performance. We think the shortest average distance of the network, $\langle d \rangle$, accounts for the experiment results. Smaller $\langle d \rangle$ means less long paths in the network, and vice versa. For example, in Table 2, the optimal AUC corresponds to a larger α, 0.62, in FW, and the smallest one, -0.75, in Power. In the twelve networks the values of $\langle d \rangle$ in FW and Power, however, are the smallest and the largest, i.e., 1.77 and 15.78 respectively. Therefore, the network with larger $\langle d \rangle$ needs smaller α to restrain the inefficient influence of long paths, and vice versa. According to the discussion above, we speculate that the value of α is associated with the shortest average distance of the network.

Besides the performance, the computational complexity of the indices is also important and discussed in the paper. According to the definitions of the indices, we get the calculation complexity of CN, AA, RA and LP as $O(N^3)$ and t-step SRW as $O(mN^3)$. N denotes the node number of network and the constant m is far less than N. Although with the same time

complexity of $O(mN^3)$ of SRW, more than CN, AA,RA and LP, SI shows better AUC performance in 8 out of 12 networks, and is the same as SRW in the rest four networks (see boldface values in Table 2). This also proves the availability of SI in different networks.

5. Conclusions

A novel index SI considering the significant influence of endpoints for link prediction in complex network is proposed in this paper. In the complex network, the connecting ability between endpoints makes up the influence of the endpoints. We assume that two-hop paths bring in the significant influence and paths with three or more hops produce the weak influence in transferring resource. Therefore, two-hop paths denoted by the common neighbors are separated from the long paths. For receiving the good prediction performance, an adjustable parameter α is set to penalize the weak influence which is derived from the long paths. To verify our assumption, experiments are implemented on twelve real-world networks and the results are compared with five traditional indices, CN, AA, RA, LP and SRW. The experiment results show that the prediction performances of classical indices are inferior for not considering the significant influence produced by the short path in the networks, especially when endpoints have many links but do not effectively transmit influence to target endpoints. From the experiment, SI efficiently overcomes the deficiency of the traditional indices and has the best prediction performance in contrast.

Acknowledgments

This research is supported in part by National Science and Technology Major Project of the Ministry of Science and Technology (2017ZX03001012-003), in part by National Natural Science Foundation of China (61461136002 and 61602048), in part by Fundamental Research Funds for the Central Universities, and in part by MOE-CMCC 1-5.

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47.
[2] S.N. Dorogovtsev, A. Goltsev, J.F.F. Mendes, Pseudofractal scale-freeweb, Phys. Rev. E 65 (6) (2002) 066122.
[3] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424 (4) (2006) 175–308.
[5] L.D.F. Costa, F.A. Rodrigues, G. Travieso, P. Villas Boas, Characterization of complex networks: A survey of measurements, Adv. Phys. 56 (1) (2007) 167–242.
[6] L. Getoor, C.P. Diehl, Link mining: a survey, ACM SIGKDD Explor. Newslett. 7 (2) (2005) 3–12.
[7] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.
[8] H. Mamitsuka, Mining from protein-protein interactions, Data Min. Knowl. Discov. 2 (5) (2012) 400–410.
[9] C.V. Cannistraci, G. Alanis-Lobato, T. Ravasi, From link-prediction in brain connectomes and protein interactomes to the local-community paradigm in complex networks, Sci. Rep. 3 (2013) 1613.
[10] J. Zhang, Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction, Info. Proc. Mgmt. 53 (1) (2017) 42–51.
[11] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Acad. Nat. Sci. 106 (52) (2009) 22073–22078.
[12] S. Scellato, A. Noulas, C. Mascolo, Exploiting place features in link prediction on location-based social networks, in: The 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 1046–1054.
[13] D. Wang, D. Pedreschi, C. Song, et al., Human mobility, social ties, and link prediction, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 1100–1108.
[14] Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, in: The 5th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM, 2005, pp. 141–142.
[15] L. Lü, M. Medo, C.H. Yeung, et al., Recommender systems, Phys. Rep. 519 (1) (2012) 1–49.
[16] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.
[17] Z. Yin, M. Gupta, T. Weninger, et al., Linkrec: a unified framework for link recommendation with user attributes and graph structure, in: The 19th International Conference on World Wide Web, ACM, 2010, pp. 1211–1212.
[18] R. Schifanella, A. Barrat, C. Cattuto, et al., Folks in folksonomies: social link prediction from shared metadata, in: The Third ACM International Conference on Web Search and Data Mining, ACM, 2010, pp. 271–280.
[19] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[20] M.E.J. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2) (2001) 025102.
[21] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1986.
[22] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1–34.
[23] E. Ravasz, A.L. Somera, D.A. Mongru, et al., Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[24] E. Leicht, P. Holme, M.E.J. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2) (2006) 026120.
[25] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Netw. 25 (3) (2003) 211–230.
[26] T. Zhou, L. Lü, Y.C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (4) (2009) 623–630.
[27] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (4) (2009) 046122.
[28] W. Liu, L. Lü, Link prediction based on local random walk, Europhys. Lett. 89 (5) (2010) 58007.
[29] X. Zhu, H. Tian, S.M. Cai, et al., Predicting missing links via significant paths, Europhys. Lett. 106 (2014) 18008.
[30] Y. Liu, M. Tang, T. Zhou, et al., Improving the accuracy of the k-shell method by removing redundant links: from a perspective of spreading dynamics, Sci. Rep. 5 (2015) 13172.
[31] S. Zeng, Link prediction based on local information considering preferential attachment, Physica A 443 (2016) 537–542.
[32] N.M. Ahmed, L. Chen, Y. Wang, et al., Sampling-based algorithm for link prediction in temporal networks, Inform. Sci. 374 (2016) 1–14.
[33] N.A. Christakis, J.H. Fowler, Social contagion theory: examining dynamic social networks and human behavior, Stat. Med. 32 (4) (2013) 556–577.

[34] V. Batagelj, A. Mrvar, Pajek-program for large network analysis, Connections 21 (2) (1998) 47–57.
[35] D. Bu, Y. Zhao, L. Cai, et al., Topological structure analysis of the protein-protein interaction network in budding yeast, Nucleic Acids Res. 31 (9) (2003) 2443–2450.
[36] J. Bascompte, Food web cohesion, Ecology 85 (2) (2004) 352–358.
[37] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (6684) (1998) 440–442.
[38] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104.
[39] R. Guimera, L. Danon, A. Diaz-Guilera, et al., Self-similar community structure in a network of human interactions, Phys. Rev. E 68 (6) (2003) 065103.
[40] P.M. Gleiser, L. Danon, Community structure in jazz, Adv. Complex Syst. 6 (04) (2003) 565–573.
[41] R. Ackland, et al. Available at http://incsub.org/blogtalk/images/robertackland.pdf.
[42] D. Van Welden, Mapping system theory problems to the field of knowledge discovery in databases, in: Proceedings of FUBUTEC'2004: 1st Future Business Technology Conference, EUROSIS, 2004, pp. 55–59.
[43] L. Isella, J. Stehlé, A. Barrat, et al., What's in a crowd? Analysis of face-to-face behavioral networks, J. Theoret. Biol. 271 (1) (2011) 166–180.
[44] N. Blagus, L. Šubelj, M. Bajec, Self-similar scaling of density in complex real-world networks, Physica A 391 (8) (2012) 2794–2802.
[45] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (20) (2002) 208701.
[46] J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148 (3) (1983) 839–843.