Hindawi
Complexity
Volume 2018, Article ID 7312603, 11 pages
https://doi.org/10.1155/2018/7312603

Research Article

Predicting Missing Links Based on a New Triangle Structure

Shenshen Bai,[1,2] Longjie Li ,[1] Jianjun Cheng,[1] Shijin Xu,[1] and Xiaoyun Chen [1]

[1]*School of Information Science & Engineering, Lanzhou University, Lanzhou 730000, China*
[2]*Department of Electronic and Information Engineering, Lanzhou Vocational Technical College, Lanzhou 730070, China*

Correspondence should be addressed to Longjie Li; ljli@lzu.edu.cn and Xiaoyun Chen; chenxy@lzu.edu.cn

Received 1 May 2018; Revised 17 October 2018; Accepted 12 November 2018; Published 2 December 2018

Guest Editor: Katarzyna Musial

With the rapid growth of various complex networks, link prediction has become increasingly important because it can discover the missing information and predict future interactions between nodes in a network. Recently, the CAR and CCLP indexes have been presented for link prediction by means of different triangle structure information. However, both indexes may lose the contributions of some shared neighbors. We propose in this work a new index to make up the weakness and then improve the accuracy of link prediction. The proposed index focuses on a new triangle structure, i.e., the triangle formed by one seed node, one common neighbor, and one other node. It emphasizes the importance of these triangles but does not ignore the contribution of any common neighbor. In addition, the proposed index adopts the theory of resource allocation by penalizing large-degree neighbors. The results of comparison with CN, AA, RA, ADP, CAR, CAA, CRA, and CCLP on 12 real-world networks show that the proposed index outperforms the compared methods in terms of AUC and ranking score.

1. Introduction

As a fundamental research hotspot in complex network analysis, link prediction has a wide range of applications in both theory and reality, such as analysis of network evolution [1, 2], recommendation system [3], and checking potential interactions between proteins in biological networks [4, 5]. The basic task of link prediction is to estimate the missing or latent existent links between unconnected nodes in a network [6, 7]. To date, a host of algorithms and models have been proposed for link prediction [6, 8, 9]. Reference [8] groups them into two ways: *similarity-based approaches* and *learning-based approaches*. A similarity-based approach computes similarity scores between unconnected nodes based on the known information. Then, a ranked list of node pairs in descending order according to their similarity scores is obtained and the node pairs at the top are thought most likely to have links. A learning-based approach formalizes the link prediction problem into a binary classification task [10] and uses machine learning methods to solve the problem. The key job in a learning-based approach is to construct the feature vectors of node pairs. In general, learning-based approaches are more complicated than similarity-based ones.

The hypothesis behind similarity-based approaches is *the more similar that two nodes are, the more likely that a link exists between them* [8]. This idea is simple and intuitive. Thus, the study of this kind of approaches has become the mainstream [6, 9]. The Common Neighbors (CN) index [11], as its name suggests, simply counts the number of common neighbors between two nodes. The Adamic-Adar (AA) [12] and Resource Allocation (RA) [13] indexes are two variants of the CN index; they penalize the contributions of large-degree common neighbors. These indexes are called local methods because they only use local structure information. Besides, some global and quasilocal methods have also been proposed by researchers, such as Katz [14], SimRank [15], Random Walks with Restart [16], Local Path [17], FriendLink [18], and Local Random Walk [19].

With the increasing growth of sizes of complex networks, local methods are still good candidates because they are more efficient in terms of running time than global and quasilocal methods. Therefore, we focus in this study on local methods. Recently, Cannistraci *et al.* proposed the CAR index [20], which suggests that links between the common neighbors, i.e., *local-community-links* (LCLs), are more valuable than common neighbors in link prediction. In CAR index, a *local*

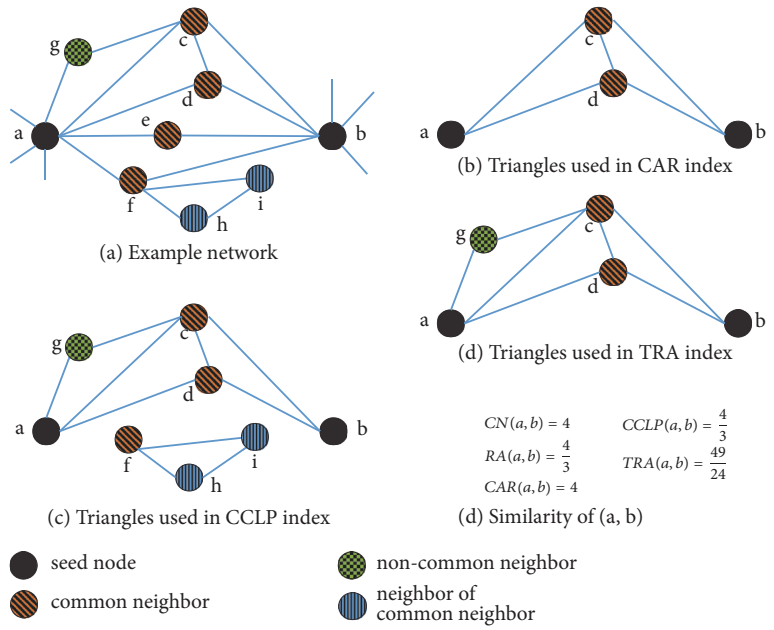

Figure 1: Triangles used in similarity indexes.

community is a triangle passing through two common neighbors and one seed node. In the example network shown in Figure 1(a), there is one LCL between the common neighbors of seed nodes a and b (see Figure 1(b)). Thus, CAR index assigns a similarity score of four to nodes a and b. However, if we remove the link between c and d, CAR will assign a zero similarity score to a and b, even though they have four common neighbors. In addition, the idea of LCL is also plugged into AA, RA, and Jaccard indexes [20]. Later, Wu *et al.* proposed the CCLP index based on the clustering coefficients of common neighbors. This index considers all triangles passing through a common neighbor. For the example network in Figure 1(a), there are triangles passing through nodes c, d, and f, respectively (see Figure 1(c)). Thus, CCLP index accumulates the clustering coefficients of nodes c, d, and f when calculating the similarity between a and b, but utterly neglects the contribution of node e. In real-world networks, it is possible that there are no triangles passing through some or even all shared neighbors of one node pair. Thus, CAR and CCLP indexes may assign a very low or even zero similarity score to the node pair, even if it has many common neighbors.

In this paper, we defines a new type of triangle structure, called TRA-triangle, which is formed by one seed node, one common neighbor, and one other node (see Figure 1(d)). Based on the TRA-triangle, a new similarity index, namely, TRA index, is proposed for link prediction. This index suggests that the common neighbors that can form TRA-triangles with a seed node are more important than others. In addition, the proposed index also penalizes the large-degree neighbors, as done in RA index [13]. Although all the TRA, CAR-based, and CCLP indexes are based on triangle structures, the intuitions behind them are different. The CAR-based indexes believe that LCLs are more valuable than common neighbors. The CCLP index is inspired by CAR index but employs all triangles passing through common neighbors, while the TRA index, which only uses the TRA-triangles, strikes a balance between CAR and CCLP. Furthermore, as aforementioned, CAR-based and CCLP indexes lose the contribution of those common neighbors with no triangles passing through them, whereas TRA index counts the contribution of all kinds of common neighbors. Therefore, TRA index can achieve better prediction accuracy than CAR-based indexes and CCLP index. The accuracy of TRA index is evaluated on 12 real-world networks from various fields. The experimental results show that our index is far superior to CAR-based indexes and CCLP index. Take the network of HEP as an example, which is a very sparse network, the improvements made by TRA on CAR and CCLP, under the metric of AUC, are up to 26.9% and 4.2%, respectively.

The rest of the paper is structured as follows. In Section 2, we give the description of the link prediction problem and the evaluation metrics, list the compared methods and networks, and depict the Wilcoxon signed-ranks test. Section 3 introduces the proposed method. In Section 4, the experimental results and performance analysis of the proposed method are presented. Finally, Section 5 concludes this work.

2. Preliminaries

2.1. Problem Description and Metric. Given an undirected and unweighted network $G(V, E)$, in which V and E are the node set and link set, respectively, in this study, multilinks and self-loops are not allowed. Let $N = |V|$ be the number of nodes in the network, and let U be the universal possible link set, which contains $N(N-1)/2$ possible links. Then, the set of nonobserved links or nonexisting links is $U-E$. Suppose there are some missing links in $U - E$, the task of link prediction

is to find those links. A similarity-based approach assigns a similarity score to each node pair in $U - E$ and assumes that the higher score a node pair has, the more likely there is a link between them.

To test the performance of a similarity index, we randomly divide the link set E into two parts: training set E_{tr} and testing set E_{ts}, such that $E = E_{tr} \cup E_{ts}$ and $E_{tr} \cap E_{ts} = \emptyset$. E_{tr} is supposed to be the observed information, and E_{ts} is used for testing. Two parameter-free metrics are employed to quantify the accuracy of link prediction algorithms: AUC [6] and ranking score [21, 22]. In this situation, the AUC score can be interpreted as the probability that a randomly selected missing link (i.e., a link in E_{ts}) is given a higher score than a randomly selected nonexistent link (i.e., a link in $U - E$). When implementing, if we perform n independent comparisons, there are n_1 times that the missing link has higher score and n_2 times that they have the same score. The AUC value is then computed as

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

Ranking score (RS) takes the ranks of links in testing set after sorting in descend order according to their similarity scores into consideration. Let $H = U - E_{tr}$ be the set of nonobserved links. Let e_i be a missing link in E_{ts} and r_i be its rank. The ranking score of e_i is defined as $RS(e_i) = r_i/|H|$, and the ranking score of the link prediction result is as follows:

$$RS = \frac{1}{|E_{ts}|} \sum_{e_i \in E_{ts}} RS(e_i) = \frac{1}{|E_{ts}|} \sum_{e_i \in E_{ts}} \frac{r_i}{|H|}. \tag{2}$$

Note that the AUC value is the higher the better, whereas the ranking score is the smaller the better.

2.2. Local Similarity Indexes. As yet, many similarity indexes have been proposed for link prediction [6, 8, 9]. Here, we list some local similarity indexes that will be used in our experiments for the purpose of comparison.

(1) Common Neighbor (CN) index [11] defines the similarity between x and y as the number of their common neighbors, which is

$$CN(x, y) = \left|\Gamma(x) \cap \Gamma(y)\right|, \tag{3}$$

where $\Gamma(x)$ denotes the set of neighbors of node x.

(2) Adamic-Adar (AA) index [12] is a variant of CN index, which believes that small-degree neighbors have more contributions than large-degree neighbors when computing similarity. Its definition is as follows:

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}, \tag{4}$$

where k_z is the degree of node z.

(3) Resource Allocation (RA) index [13] defines the similarity between x and y as the amount of resource that y received from x through their common neighbors, which is

$$RA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}. \tag{5}$$

(4) Adaptive Degree Penalization (ADP) index [23] penalizes a common neighbor according to its degree and the average clustering coefficient of the network. Therefore, it can automatically adapt to the network. The definition of ADP index is as follows:

$$ADP(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} k_z^{-\beta C}, \tag{6}$$

where β is a constant and C is the average clustering coefficient of the network. We set $\beta = 2.5$, as suggested by the authors.

(5) CAR index [20] suggests that two seed nodes are more likely to link together if there are links between their common neighbors, which is defined as

$$CAR(x, y) = CN(x, y) \cdot \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{L(z)}{2}, \tag{7}$$

where $L(z)$ is the number of links between z and other common neighbors of x and y.

(6) CAA and CRA indexes [20] are generated by plugging the idea of CAR index into the AA and RA indexes, respectively, which are defined as

$$CAA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{L(z)}{\log k_z}, \tag{8}$$

$$CRA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{L(z)}{k_z}. \tag{9}$$

(7) CCLP index [24] computes the similarity between x and y by employing clustering coefficient of common neighbors, which is

$$CCLP(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} CC_z, \tag{10}$$

where CC_z denotes the clustering coefficient of node z, which is

$$CC_z = \frac{2t_z}{k_z(k_z - 1)}, \tag{11}$$

in which t_z is the number of triangles passing through node z.

2.3. Networks. In this study, we use 12 real-world networks drawn from various fields to evaluate the effectiveness of link prediction methods.

(1) Advogato (ADV): a social network whose users are mainly free and open source software developers [25].

(2) C.elegans (CE): the neural network of a Caenorhabditis elegans worm [26].

(3) Dolphin: a social network of 62 dolphins in a community living off Doubtful Sound, New Zealand [27].

(4) Email: a network of email interchanges between members of a university [28].

(5) Foodweb (FW): a food web in Florida Bay during the rainy season [29].

(6) Hamster: a friendship network between users on hamsterster.com [30].

(7) HEP: the coauthorships network of scientists who posted preprints on the high-energy theory archive from 1995 to 1999 [31].

(8) Karate: the social network of a karate club at a US university [32].

(9) Political blogs (PB): a network of blogs about US politics [33].

(10) USAir: a network of the US air transportation system [6].

(11) Word: an adjacency network of common adjectives and noun in the novel "David Copperfield" by Charles Dickens [34].

(12) Yeast: the protein-protein interaction network of budding yeast [35].

In this work, all the aforementioned networks are treated as undirected and unweighted networks, and only the giant component of each network is used. Table 1 lists the basic statistics of the giant components of these networks.

Given network $G(V, E)$, suppose x, y be two seed nodes. (x, y) is called *a seed node pair with common neighbors* if they have at least one common neighbor. P_Λ denotes the set of seed node pairs with common neighbors, formally

$$P_\Lambda = \{(x, y) \mid (x, y) \notin E \wedge \Gamma(x) \cap \Gamma(y) \neq \emptyset\}. \tag{12}$$

Let x, y be two seed nodes, and z is one of their common neighbors. If $CC_z = 0$, we call z is a *zero-triangle-neighbor*; otherwise, z is a *triangle-neighbor*. If $L(z) \neq 0$, z is called a *CAR-triangle-neighbor* and if $\triangle(x, y; z) \neq 0$ (see (18)), z is called a *TRA-triangle-neighbor*. Let $S_\triangle$ be the set of triangle-neighbors, and S_{CAR}, S_{TRA} denote the sets of CAR- and TRA-triangle-neighbors, respectively. Clearly, $S_{CAR} \subseteq S_{TRA} \subseteq S_\triangle$. Let $P_\exists(\overline{S}_\triangle)$ and $P_\forall(\overline{S}_\triangle)$ be two subsets of P_Λ. For any pair in $P_\exists(\overline{S}_\triangle)$, at least one of their shared neighbors is not a triangle-neighbor, and for any pair in $P_\forall(\overline{S}_\triangle)$, all of their shared neighbors are not triangle-neighbors. More explicitly,

$$\begin{aligned} &P_\exists\left(\overline{S}_\triangle\right) \\ &\quad = \{(x, y) \in P_\Lambda \mid \exists z \in \Gamma(x) \cap \Gamma(y) \wedge z \notin S_\triangle\}, \\ &P_\forall\left(\overline{S}_\triangle\right) \\ &\quad = \{(x, y) \in P_\Lambda \mid \forall z \in \Gamma(x) \cap \Gamma(y) \wedge z \notin S_\triangle\}. \end{aligned} \tag{13}$$

Similarly, we define $P_\exists(\overline{S}_{TRA})$, $P_\forall(\overline{S}_{TRA})$, $P_\exists(\overline{S}_{CAR})$, and $P_\forall(\overline{S}_{CAR})$, which are

$$\begin{aligned} &P_\exists\left(\overline{S}_{TRA}\right) \\ &\quad = \{(x, y) \in P_\Lambda \mid \exists z \in \Gamma(x) \cap \Gamma(y) \wedge z \notin S_{TRA}\}, \\ &P_\forall\left(\overline{S}_{TRA}\right) \\ &\quad = \{(x, y) \in P_\Lambda \mid \forall z \in \Gamma(x) \cap \Gamma(y) \wedge z \notin S_{TRA}\}, \\ &P_\exists\left(\overline{S}_{CAR}\right) \\ &\quad = \{(x, y) \in P_\Lambda \mid \exists z \in \Gamma(x) \cap \Gamma(y) \wedge z \notin S_{CAR}\}, \\ &P_\forall\left(\overline{S}_{CAR}\right) \\ &\quad = \{(x, y) \in P_\Lambda \mid \forall z \in \Gamma(x) \cap \Gamma(y) \wedge z \notin S_{CAR}\}. \end{aligned} \tag{14}$$

Correspondingly, the ratios of those subsets to P_Λ are, respectively, defined as

$$\begin{aligned} R_\exists\left(\overline{S}_\triangle\right) &= \frac{\left|P_\exists\left(\overline{S}_\triangle\right)\right|}{|P_\Lambda|}, \\ R_\forall\left(\overline{S}_\triangle\right) &= \frac{\left|P_\forall\left(\overline{S}_\triangle\right)\right|}{|P_\Lambda|}, \\ R_\exists\left(\overline{S}_{TRA}\right) &= \frac{\left|P_\exists\left(\overline{S}_{TRA}\right)\right|}{|P_\Lambda|}, \\ R_\forall\left(\overline{S}_{TRA}\right) &= \frac{\left|P_\forall\left(\overline{S}_{TRA}\right)\right|}{|P_\Lambda|}, \\ R_\exists\left(\overline{S}_{CAR}\right) &= \frac{\left|P_\exists\left(\overline{S}_{CAR}\right)\right|}{|P_\Lambda|}, \\ R_\forall\left(\overline{S}_{CAR}\right) &= \frac{\left|P_\forall\left(\overline{S}_{CAR}\right)\right|}{|P_\Lambda|}. \end{aligned} \tag{15}$$

Table 2 lists these ratios over the 12 networks.

2.4. Wilcoxon Signed-Ranks Test. The Wilcoxon signed-ranks test is a nonparametric statistical hypothesis test used to check whether two methods perform equally well over multiple networks [38, 39]. Let d_i be the difference in performance scores of two link prediction methods on the ith network. The differences are ranked in accordance with their absolute values; in case of ties, average ranks are assigned. Let R^+ be the sum of ranks for the networks on which the second method outperformed the first, and R^- the sum of ranks for the opposite. For a larger number of networks, the statistics

$$z = \frac{T - (1/4) N (N + 1)}{\sqrt{(1/24) N (N + 1) (2N + 1)}} \tag{16}$$

is distributed approximately normally [39]. In (16), $T = \min(R^+, R^-)$ and N is the number of networks.

With $\alpha = 0.05$, if z is small than -1.96, we reject the null-hypothesis, which states that both methods perform equally well.

Table 1: The basic structural features of the giant components of the 12 networks. $|V|$ and $|E|$ are the total numbers of nodes and edges, respectively. D denotes the density, which is $D = 2|E|/|V|(|V|-1)$. $\langle k\rangle$ and $\langle d\rangle$ present the average degree and the average shortest distance, respectively. C and r indicate the clustering coefficient [26] and assortative coefficient [36], respectively. H is the degree heterogeneity [6], defined as $H = \langle k^2\rangle/\langle k\rangle^2$, and e is the network efficiency [37].

Networks	$\|V\|$	$\|E\|$	D	$\langle k\rangle$	$\langle d\rangle$	C	r	H	e
ADV	5042	39227	3.1E-03	15.560	3.275	0.253	-0.096	5.303	0.324
CE	297	2148	4.9E-02	14.465	2.455	0.292	-0.163	1.801	0.445
Dolphin	62	159	8.4E-02	5.129	3.357	0.259	-0.044	1.327	0.379
Email	1133	5451	8.5E-03	9.622	3.606	0.220	0.078	1.942	0.300
FW	128	2075	2.6E-01	32.422	1.776	0.335	-0.112	1.237	0.622
Hamster	1788	12476	7.8E-03	13.955	3.453	0.143	-0.089	3.264	0.317
HEP	5835	13815	8.1E-04	4.735	7.026	0.506	0.185	1.926	0.155
Karate	34	78	1.4E-01	4.588	2.408	0.571	-0.476	1.693	0.492
PB	1222	16714	2.2E-02	27.355	2.738	0.320	-0.221	2.971	0.398
USAir	332	2126	3.9E-02	12.807	2.738	0.625	-0.208	3.464	0.406
Word	112	425	6.8E-02	7.589	2.536	0.173	-0.129	1.815	0.442
Yeast	2224	6609	2.7E-03	5.943	4.376	0.138	-0.105	2.803	0.246

Table 2: The ratios of various seed pairs over the 12 networks.

Networks	$R_\exists(\overline{S}_\triangle)$	$R_\forall(\overline{S}_\triangle)$	$R_\exists(\overline{S}_{CAR})$	$R_\forall(\overline{S}_{CAR})$	$R_\exists(\overline{S}_{TRA})$	$R_\forall(\overline{S}_{TRA})$
ADV	0.001	0.001	0.807	0.750	0.018	0.014
CE	0.001	0.000	0.768	0.670	0.012	0.008
Dolphin	0.020	0.011	0.857	0.817	0.089	0.069
Email	0.008	0.005	0.881	0.841	0.070	0.057
FW	0.000	0.000	0.240	0.136	0.005	0.000
Hamster	0.014	0.005	0.893	0.817	0.074	0.048
HEP	0.007	0.007	0.881	0.876	0.029	0.028
Karate	0.004	0.000	0.743	0.706	0.034	0.011
PB	0.001	0.000	0.577	0.497	0.010	0.007
USAir	0.000	0.000	0.510	0.509	0.005	0.005
Word	0.067	0.028	0.799	0.735	0.108	0.062
Yeast	0.083	0.063	0.945	0.931	0.357	0.324

3. The Proposed Index

The link prediction problem has a familiar relationship with the network evolving mechanism [2, 40]. A recently proposed triangle growth mechanism demonstrates that various key features observed in most real-world networks can be generated in simulated networks [41]. Therefore, triangle structure information has an important effect in link formation.

In this work, we focus on a new triangle structure, namely *TRA-triangle*. A TRA-triangle passes through one seed node, one common neighbor, and one other node. In our opinion, the common neighbors that can form TRA-triangles are more important than others. Given two nodes u and v, we denote the number of triangles passing through them as $\triangle(u, v)$, which is

$$\triangle(u,v) = \begin{cases} CN(u,v), & \text{if } (u,v) \in E \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

For the example network in Figure 1(a), the triangles used for seed nodes a, b are shown in Figure 1(d). Clearly, $\triangle(a,c) = 2$ and $\triangle(a,d) = 1$. Thus, node c is in more close contact with a than d. Given seed nodes x and y, z is one of their common neighbors. Function $\triangle(x, y; z)$ sums up the number of TRA-triangles formed by x, z, and y, z, which is

$$\triangle(x, y; z) = \triangle(x, z) + \triangle(y, z). \tag{18}$$

In this paper, we propose a new similarity index, by combining the aforementioned triangle structure and the idea of RA index [13]. For the convenience of statement, we name our new method *TRA* index. Its definition is

$$TRA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1 + \triangle(x, y; z)/2}{k_z}. \tag{19}$$

In (19), the numerator is $1 + \triangle(x, y; z)/2$. Therefore, the TRA index does not miss the effect of any common neighbor. If all common neighbors are zero-triangle-neighbors, TRA degenerates to RA. For the example network in Figure 1(a), $TRA(a, b) = (1+3/2)/4+(1+2/2)/3+(1+0/2)/2+(1+0/2)/4 = 49/24$.

Table 3: The AUC of different methods in 12 networks. The results are the average of 50 independent implementations with $|E_{ts}|/|E| = 0.1$. The best performance for each network is emphasized by boldface.

	CN	AA	RA	ADP	CAR	CAA	CRA	CCLP	TRA
ADV	0.8992	0.9026	0.9030	0.9033	0.8054	0.8051	0.8063	0.9011	**0.9043**
CE	0.8450	0.8613	0.8662	0.8654	0.7657	0.7677	0.7704	0.8625	**0.8713**
Dolphin	0.7832	0.7863	0.7854	**0.7866**	0.6475	0.6473	0.6473	0.7804	0.7850
Email	0.8471	0.8491	0.8488	0.8491	0.6994	0.6995	0.6996	0.8452	**0.8493**
FW	0.6053	0.6071	0.6114	0.6097	0.6192	0.6271	0.6321	0.6323	**0.6890**
Hamster	0.8037	0.8067	0.8074	0.8067	0.6542	0.6542	0.6543	0.8075	**0.8127**
HEP	0.8984	**0.8987**	**0.8987**	**0.8987**	0.7079	0.7079	0.7079	0.8624	0.8985
Karate	0.6985	0.7409	0.7523	0.7532	0.5848	0.5881	0.5880	0.7085	**0.7755**
PB	0.9192	0.9226	0.9239	0.9242	0.8926	0.8929	0.8946	0.9217	**0.9282**
USAir	0.9357	0.9466	0.9522	**0.9523**	0.9136	0.9158	0.9202	0.9391	0.9452
Word	0.6656	0.6649	0.6621	0.6651	0.5717	0.5713	0.5714	0.6727	**0.6809**
Yeast	0.7041	0.7047	0.7045	0.7047	0.5994	0.5994	0.5994	0.6972	**0.7054**

4. Experimental Results

Table 3 lists the predicted results of different methods in terms of AUC on the 12 networks. The results are obtained by averaging over 50 independent realizations for each network with testing set containing 10% links. The highest AUC value for each network is highlighted in boldface. Clearly, TRA index gets nine best results over the 12 networks. Meanwhile, TRA index outperforms the CAR, CAA, CRA, and CCLP indexes on all networks. We can see from Table 2 that, on most of the networks, there exist varying degrees of such seed node pairs with common neighbors that belong to $P_{\exists}(\overline{S}_{\triangle})$ and/or $P_{\forall}(\overline{S}_{\triangle})$. As stated in Introduction, CCLP index will give lower or zero similarity scores to those pairs. Furthermore, both values of $R_{\exists}(\overline{S}_{CAR})$ and $R_{\forall}(\overline{S}_{CAR})$ are very high on most of the networks. Particularly, on Dolphin, Email, Hamster, HEP, and Yeast, the corresponding values of $R_{\forall}(\overline{S}_{CAR})$ are greater than 0.8. This phenomenon indicates that only a very small fraction of seed node pairs with common neighbors on those networks can be assigned similarity scores by CAR-based indexes. Although there are some seed node pairs belonging to $P_{\exists}(\overline{S}_{TRA})$ and/or $P_{\forall}(\overline{S}_{TRA})$, TRA index still can assign reasonable similarity scores to them. Therefore, the results of TRA index in Table 3 are better than them of CAR, CAA, CRA, and CCLP indexes. For CN, AA, RA, and ADP indexes, ADP index performs the best, since it can penalize common neighbors by automatically adapting to the network. On Dolphin, HEP, and USAir, ADP index obtains the best accuracy; the performance of our index approximates to the best. In addition, TRA index achieves much better AUC scores than others on FW and Karate. This result suggests that TRA-triangles play an important role on these two networks. From Table 1, both networks are dense ones. Roughly speaking, the probability that there exist TRA-triangle-neighbors between seed nodes on dense networks is more than on sparse ones.

To check whether the proposed index is significantly different with compared methods, we applied Wilcoxon signed-ranks test [39] based on the results in Table 3. The pairwise test results are presented in Figure 2. From the statistical point of view, our index is significantly better than others except ADP index, because ADP index has the capability of adapting to the structure of a network automatically. Although there is no statistical difference between our index and ADP index according to Wilcoxon signed-ranks test, our index performs better than ADP index in terms of AUC.

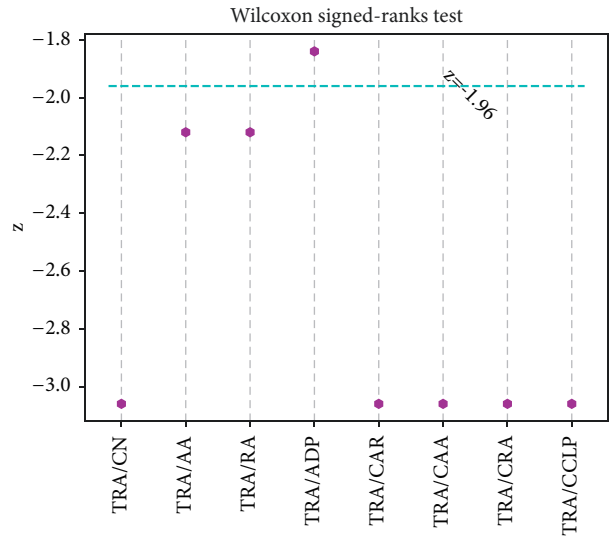

Figure 2: The results of Wilcoxon signed-ranks test based on Table 3. With $\alpha = 0.05$, if $z <= -1.96$, the null-hypothesis is rejected.

Figure 3 exhibits the changes of AUC on 12 networks when the proportion of E_{ts} in E increases from 10% to 20%. It is quite evident from Figure 3 that the AUC values of all indexes show downward trends when the proportion increases from 10% to 20% except on FW. The reason is that the increase of E_{ts} will decrease the size of training set E_{tr} and then will result in the number of common neighbors between seed nodes becoming small. Consequently, the difficulty of link prediction will enhance. The FW network, which possesses high average degree, small average shortest distance, and small-degree heterogeneity, is a very dense

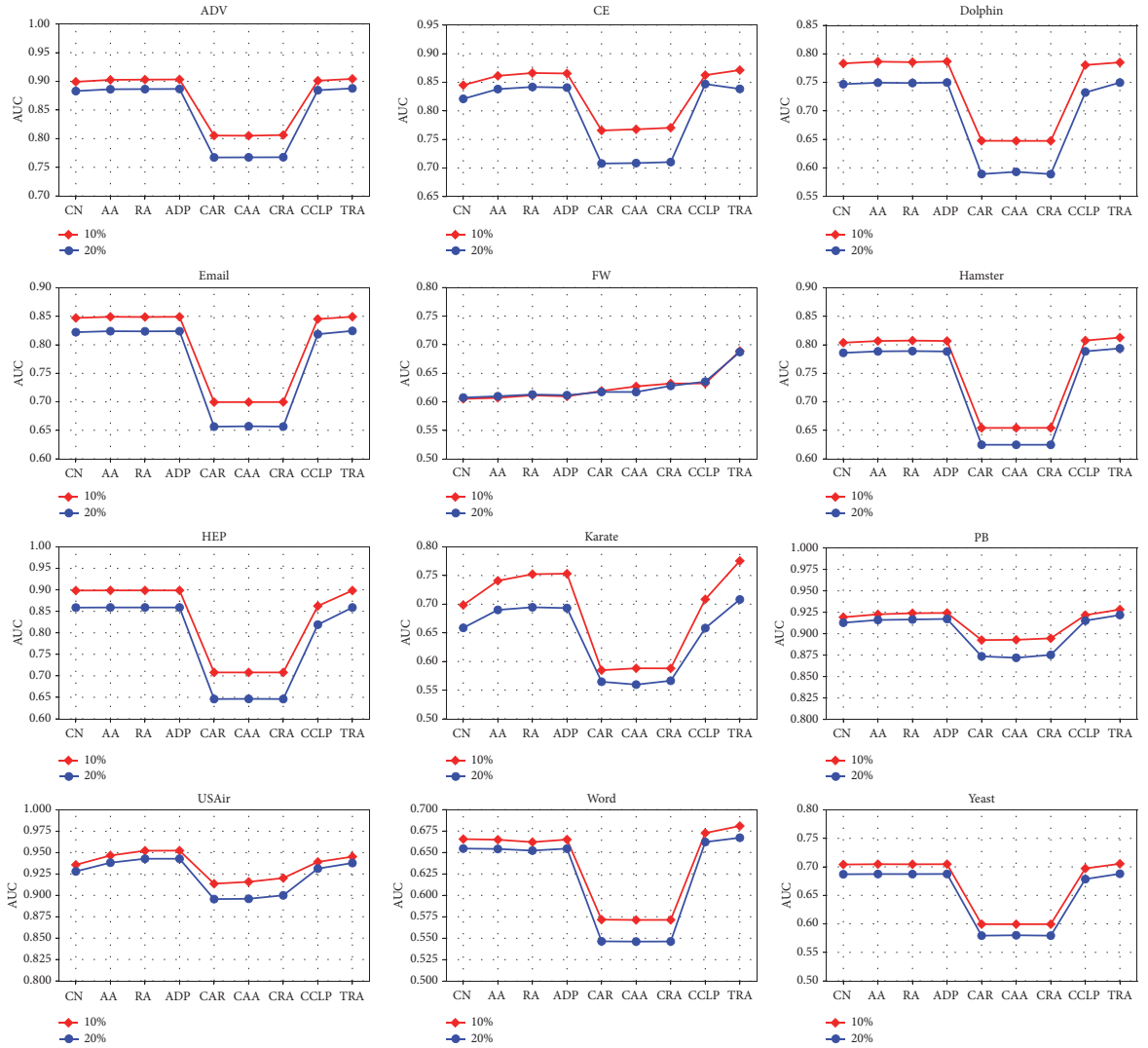

Figure 3: The changes of AUC when $|E_{ts}|/|E|$ increases from 10% to 20% on 12 networks. Each point is obtained by averaging over 50 independent realizations.

network. Therefore, the decrease of training set gives slight influence of accuracy on FW. In addition, we can observe from Figure 3 that the performance presented by all indexes on ADV, CE, Dolphin, Email, Hamster, HEP, Karate, Word, and Yeast is very similar. On these nine networks, the AUC values of CAR-based indexes are obvious lower than those of others. On the network of FW, the results of CAR-based indexes are better than those of CN, AA, RA, and ADP indexes, because FW is a very dense network in which the ratio of CAR-triangle-neighbor is very high (see Table 2). On PB and USAir, the performance of CAR-based indexes is not as bad as on other nine networks. The reason is both networks have high average degrees, small average shortest distances, and high ratio of CAR-triangle-neighbors.

Furthermore, we list the AUC values of different methods on the 12 networks when $|E_{ts}|/|E| = 0.2$ in Table 4. The results of our index outperform others on eight among the 12 networks, while CCLP index achieves the highest value on CE.

Table 5 gives the results in terms of ranking score. These results are similar to those in Table 3. The ranking score of TRA index outperforms others except on Dolphin, HEP, and USAir. The pairwise Wilcoxon signed-ranks test results are shown in Figure 4. Similar to the test in Figure 2, TRA index is significantly better than compared methods except ADP index. As depicted above, ADP has the adaptive capability and hence performs better than other compared methods.

Figure 5 describes the changes of ranking score on 12 networks when $|E_{ts}|/|E|$ increases from 10% to 20%. Clearly, all indexes yield higher ranking scores with the increase of E_{ts}. Do not forget that higher ranking score means lower accuracy. As analyzed above, FW is very dense. Thus, the changes of AUC on FW are very slight (see Figure 3). However, the changes of ranking score on FW are more

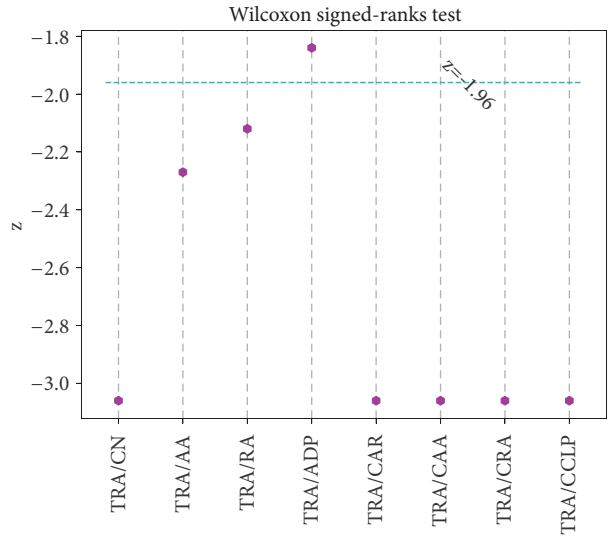

Figure 4: The results of Wilcoxon signed-ranks test based on Table 5. With $\alpha = 0.05$, if $z <= -1.96$, the null-hypothesis is rejected.

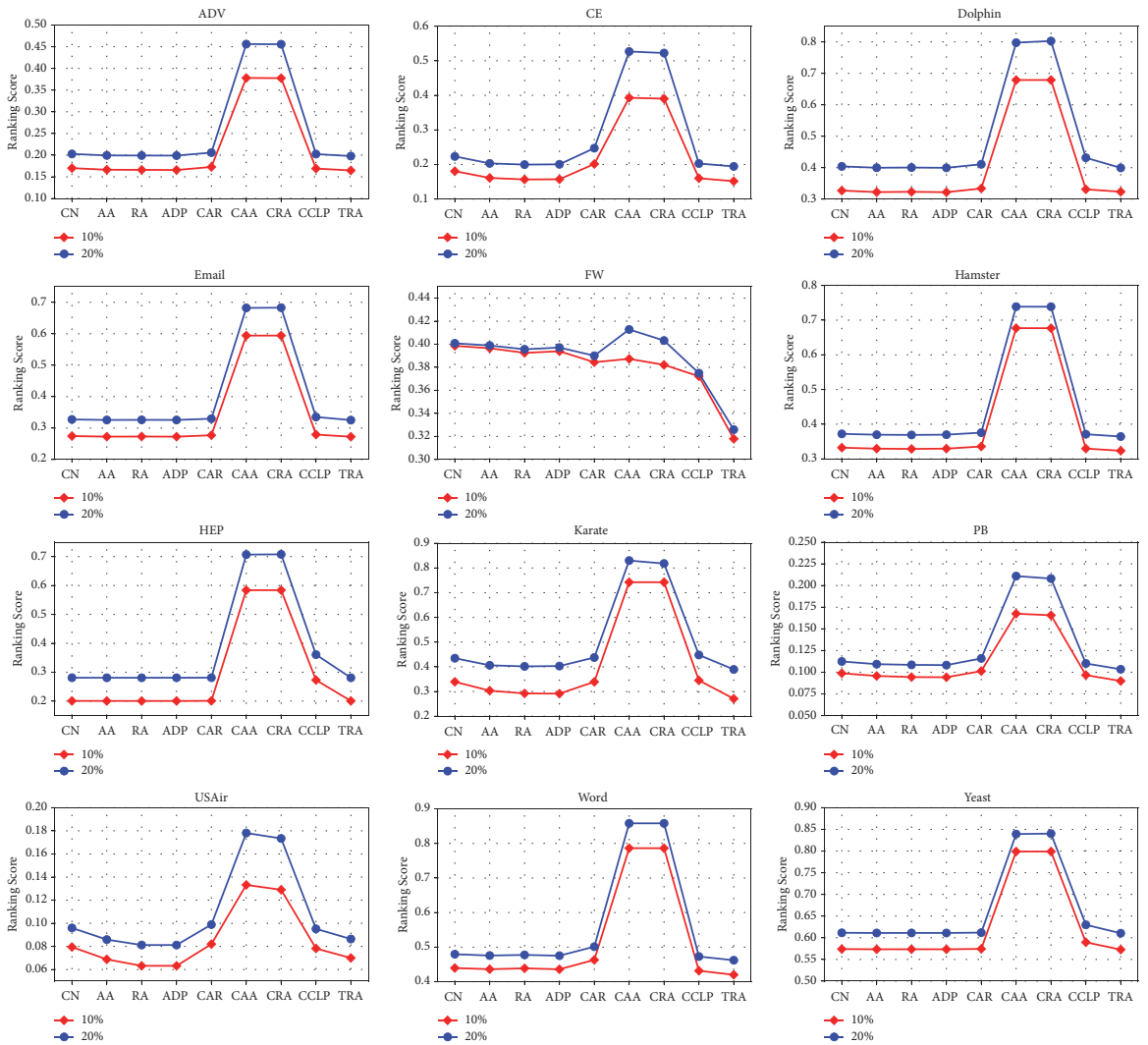

Figure 5: The changes of ranking score when $|E_{ts}|/|E|$ increases from 10% to 20% on 12 networks. Each point is obtained by averaging over 50 independent realizations.

Table 4: The AUC of different methods in 12 networks. The results are the average of 50 independent implementations with $|E_{ts}|/|E| = 0.2$. The best performance for each network is emphasized by boldface.

	CN	AA	RA	ADP	CAR	CAA	CRA	CCLP	TRA
ADV	0.8830	0.8862	0.8864	0.8867	0.7672	0.7674	0.7676	0.8845	**0.8877**
CE	0.8210	0.8381	0.8418	0.8407	0.7079	0.7086	0.7104	**0.8469**	0.8384
Dolphin	0.7468	0.7494	0.7488	**0.7497**	0.5891	0.5930	0.5890	0.7322	0.7496
Email	0.8221	0.8239	0.8236	0.8239	0.6564	0.6571	0.6565	0.8186	**0.8244**
FW	0.6075	0.6099	0.6129	0.6117	0.6174	0.6173	0.6280	0.6354	**0.6872**
Hamster	0.7859	0.7885	0.7890	0.7883	0.6246	0.6246	0.6246	0.7885	**0.7937**
HEP	0.8587	**0.8590**	**0.8590**	**0.8590**	0.6460	0.6464	0.6460	0.8190	0.8589
Karate	0.6587	0.6900	0.6946	0.6932	0.5647	0.5597	0.5664	0.6582	**0.7082**
PB	0.9128	0.9160	0.9166	0.9172	0.8736	0.8719	0.8753	0.9152	**0.9218**
USAir	0.9280	0.9382	**0.9428**	**0.9428**	0.8956	0.8960	0.9000	0.9313	0.9378
Word	0.6546	0.6541	0.6522	0.6545	0.5464	0.5461	0.5462	0.6622	**0.6672**
Yeast	0.6870	0.6874	0.6873	0.6875	0.5793	0.5800	0.5793	0.6785	**0.6879**

Table 5: The ranking score of different methods in 12 networks. The results are the average of 50 independent implementations with $|E_{ts}|/|E| = 0.1$. The best performance for each network is emphasized by boldface.

	CN	AA	RA	ADP	CAR	CAA	CRA	CCLP	TRA
ADV	0.1700	0.1663	0.1660	0.1657	0.1727	0.3780	0.3776	0.1690	**0.1647**
CE	0.1807	0.1613	0.1568	0.1574	0.2015	0.3930	0.3904	0.1603	**0.1518**
Dolphin	0.3271	0.3223	0.3232	**0.3220**	0.3338	0.6788	0.6788	0.3311	0.3234
Email	0.2745	0.2727	0.2731	0.2726	0.2771	0.5941	0.5942	0.2793	**0.2724**
FW	0.3986	0.3965	0.3925	0.3939	0.3844	0.3873	0.3821	0.3722	**0.3179**
Hamster	0.3323	0.3295	0.3287	0.3295	0.3357	0.6769	0.6768	0.3297	**0.3234**
HEP	0.2008	**0.2005**	**0.2005**	**0.2005**	0.2009	0.5839	0.5839	0.2729	0.2007
Karate	0.3393	0.3034	0.2922	0.2913	0.3391	0.7424	0.7424	0.3450	**0.2708**
PB	0.0988	0.0957	0.0944	0.0942	0.1013	0.1675	0.1657	0.0967	**0.0899**
USAir	0.0795	0.0688	**0.0632**	**0.0632**	0.0820	0.1332	0.1290	0.0782	0.0700
Word	0.4396	0.4362	0.4387	0.4360	0.4631	0.7862	0.7860	0.4317	**0.4201**
Yeast	0.5739	0.5734	0.5735	0.5734	0.5743	0.7989	0.7989	0.5895	**0.5727**

evident, especially for CAA and CRA indexes. The reason is that the calculation of ranking score considers all missing links. In addition, as seen in Figure 5, CAA and CRA indexes perform worse than CAR index according to ranking score. From the definitions of these three indexes, we find that both CAA and CRA indexes can get more negative impact than CAR index from zero-triangle-neighbors.

Finally, the ranking scores of all methods on the 12 networks with $|E_{ts}|/|E| = 0.2$ are listed in Table 6. Our index outperforms all other indexes except on HEP and USAir in terms of ranking score. These results are consistent with them of AUC. In contrast with that on FW, the influence of TRA-triangles on HEP and USAir is small.

From the above results, we can conclude that TRA index is superior to CAR-based indexes and CCLP index and performs better than common-neighbor-based methods on most of networks.

5. Conclusion and Discussion

Link prediction is an important research topic of complex network analysis and has a wide range of applications in various fields. Inspired by the triangle growth mechanism in network evolving [41], this paper proposed the TRA index for link prediction. When computing the similarity between two seed nodes, the proposed index not only counts the contributions of all common neighbors but also emphasizes the importance of the neighbors that can form TRA-triangles. To some extent, TRA-triangles reflect the close relationships between neighbors and seed nodes. In addition, the proposed index also adopts the theory of resource allocation [13] due to its effectiveness.

The accuracy of the TRA index is experimentally evaluated over 12 real-world networks from various fields in terms of AUC and ranking score. The experimental results show that the proposed index performs far better than CAR-based indexes. Meanwhile, our index outperforms the CCLP index because of the superior strategy in our index. For common-neighbor-based methods, the proposed index yields some improvements of accuracy on most of networks. These results indicate that combining the information of TRA-triangles and the theory of resource allocation in similarity index is a helpful idea for link prediction.

Table 6: The ranking score of different methods in 12 networks. The results are the average of 50 independent implementations with $|E_{ts}|/|E| = 0.2$. The best performance for each network is emphasized by boldface.

	CN	AA	RA	ADP	CAR	CAA	CRA	CCLP	TRA
ADV	0.2027	0.1993	0.1991	0.1989	0.2058	0.4561	0.4558	0.2024	**0.1978**
CE	0.2234	0.2033	0.1998	0.2006	0.2473	0.5268	0.5224	0.2029	**0.1945**
Dolphin	0.4040	0.3998	0.4004	0.3995	0.4108	0.7976	0.8029	0.4315	**0.3993**
Email	0.3274	0.3257	0.3261	0.3257	0.3298	0.6821	0.6829	0.3352	**0.3253**
FW	0.4008	0.3989	0.3956	0.3971	0.3900	0.4128	0.4032	0.3749	**0.3259**
Hamster	0.3721	0.3696	0.3690	0.3697	0.3756	0.7388	0.7387	0.3710	**0.3644**
HEP	0.2810	**0.2808**	**0.2808**	**0.2808**	0.2811	0.7070	0.7079	0.3609	0.2809
Karate	0.4347	0.4061	0.4017	0.4030	0.4376	0.8301	0.8181	0.4478	**0.3893**
PB	0.1124	0.1093	0.1085	0.1082	0.1159	0.2109	0.2080	0.1101	**0.1034**
USAir	0.0960	0.0859	**0.0812**	**0.0812**	0.0989	0.1781	0.1734	0.0952	0.0864
Word	0.4796	0.4757	0.4775	0.4753	0.5013	0.8580	0.8580	0.4728	**0.4621**
Yeast	0.6114	0.6110	0.6111	0.6110	0.6118	0.8392	0.8402	0.6298	**0.6104**

There are some improved studies for our index in future. One of them is to analyze the degree of influence of TRA-triangles on different networks and further to be adaptive to set the weight of TRA-triangles on different networks. The second is to study the application of TRA index on other topics, such as community detection and anomaly detection. In addition, for learning-based link prediction approaches, TRA index can be used as a feature for a node pair.

Data Availability

The networks used in this study are available from http://deim.urv.cat/~alexandre.arenas/data/welcome.htm, http://www-personal.umich.edu/~mejn/netdata/, http://vlado.fmf.uni-lj.si/pub/networks/data/, http://noesis.ikor.org/datasets/link-prediction, and http://konect.uni-koblenz.de/networks/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61602225) and the Fundamental Research Funds for the Central Universities (no. lzujbky-2017-192).

References

[1] Q.-M. Zhang, L. Lü, W.-Q. Wang, Y.-X. Zhu, and T. Zhou, "Potential theory for directed networks," *PLoS ONE*, vol. 8, no. 2, Article ID e55437, 2013.

[2] Q. Zhang, X. Xu, Y. Zhu, and T. Zhou, "Measuring multiple evolution mechanisms of complex networks," *Scientific Reports*, vol. 5, no. 1, 2015.

[3] L. Lü, M. Medo, C. H. Yeung, Y. Zhang, Z. Zhang, and T. Zhou, "Recommender systems," *Physics Reports*, vol. 519, no. 1, pp. 1–49, 2012.

[4] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 106, no. 52, pp. 22073–22078, 2009.

[5] S. S. Bhowmick and B. S. Seah, "Clustering and Summarizing Protein-Protein Interaction Networks: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 638–658, 2016.

[6] L. Lü and T. Zhou, "Link prediction in complex networks: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[7] L. Li, L. Qian, X. Wang, S. Luo, and X. Chen, "Accurate similarity index based on activity and connectivity of node for link prediction," *International Journal of Modern Physics B*, vol. 29, no. 17, 1550108, 15 pages, 2015.

[8] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," *Science China Information Sciences*, vol. 58, no. 1, pp. 1–38, 2014.

[9] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys*, vol. 49, no. 4, pp. 69:1–69:33, 2016.

[10] C. Ahmed, A. ElKorany, and R. Bahgat, "A supervised learning approach to link prediction in Twitter," *Social Network Analysis and Mining*, vol. 6, no. 1, 2016.

[11] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the Association for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

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

[13] 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.

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

[15] G. Jeh and J. Widom, "SimRank," in *Proceedings of the the eighth ACM SIGKDD international conference*, p. 538, Edmonton, Alberta, Canada, July 2002.

[16] H. Tong, C. Faloutsos, and J. Pan, "Fast random walk with restart and its applications," in *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*, pp. 613–622, December 2006.

[17] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, no. 4, Article ID 046122, 2009.

[18] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems," *The Journal of Systems and Software*, vol. 85, no. 9, pp. 2119–2132, 2012.

[19] W. Liu and L. Lu, "Link prediction based on local random walk," *EPL (Europhysics Letters)*, vol. 89, no. 5, Article ID 58007, 2010.

[20] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks," *Scientific Reports*, vol. 3, article 1613, no. 4, 2013.

[21] B. Chen and L. Chen, "A link prediction algorithm based on ant colony optimization," *Applied Intelligence*, vol. 41, no. 3, pp. 694–708, 2014.

[22] D. Caiyan, L. Chen, and B. Li, "Link prediction in complex network based on modularity," *Soft Computing*, vol. 21, no. 15, pp. 4197–4214, 2017.

[23] V. Martnez, F. Berzal, and J.-C. Cubero, "Adaptive degree penalization for link prediction," *Journal of Computational Science*, vol. 13, pp. 1–9, 2016.

[24] Z. Wu, Y. Lin, J. Wang, and S. Gregory, "Link prediction with node clustering coefficient," *Physica A: Statistical Mechanics and its Applications*, vol. 452, pp. 1–8, 2016.

[25] P. Massa, M. Salvetti, and D. Tomasoni, "Bowling alone and trust decline in social network sites," in *Proceedings of the 8th IEEE International Symposium on Dependable, Autonomic and Secure Computing, DASC 2009*, pp. 658–663, China, December 2009.

[26] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[27] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait?" *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

[28] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 68, no. 6, Article ID 065103, 2003.

[29] R. E. Ulanowicz and D. L. DeAngelis, "Network analysis of trophic dynamics in south florida ecosystems," in *US Geological Survey Program on the South Florida Ecosystem*, vol. 114, 45 edition, 2005.

[30] J. Kunegis, "KONECT—the koblenz network collection," in *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, pp. 1343–1350, May 2013.

[31] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.

[32] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[33] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. Election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD '05)*, pp. 36–43, ACM, 2005.

[34] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 74, no. 3, Article ID 036104, 19 pages, 2006.

[35] D. Bu, Y. Zhao, L. Cai et al., "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[36] M. E. Newman, "Mixing patterns in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 67, no. 2, 2003.

[37] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical Review Letters*, vol. 87, no. 19, Article ID 198701, 2001.

[38] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[39] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[40] W.-Q. Wang, Q.-M. Zhang, and T. Zhou, "Evaluating network models: a likelihood analysis," *EPL (Europhysics Letters)*, vol. 98, no. 2, Article ID 28004, 2012.

[41] Z. Wu, G. Menichetti, C. Rahmede, and G. Bianconi, "Emergent complex network geometry," *Scientific Reports*, vol. 5, 2015.