IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 26, 2019, accepted February 3, 2020, date of publication February 6, 2020, date of current version February 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972072

Link Prediction in Directed Networks Utilizing the Role of Reciprocal Links

JINSONG LI[1,2], JIANHUA PENG[1], SHUXIN LIU[1], XINSHENG JI[1,2], XING LI[1], AND XINXIN HU[1,3]

[1]National Digital Switching System Engineering and Technological Research Center, Zhengzhou 450001, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3]Department of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Shuxin Liu (liushuxin11@126.com)

This work was supported in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61521003, and in part by the National Natural Science Foundation of China under Grant 61803384.

ABSTRACT Link prediction in directed networks has always been a hot topic in many fields including network science, information system and data mining. Intuitively, once links are endowed with certain orientations, their reciprocate nature can potentially provide extra information for guiding link prediction. However, the role of reciprocal links in the formation of directed closure triads and their ability to enhance link prediction accuracy are not thoroughly investigated yet in existing works. In this paper, we first design an empirical test to investigate the role of reciprocal links in different types of directed networks by taking advantage of null models. Subsequently, based on solid evidence of the empirical test, two novel weighting mechanisms for link prediction are proposed utilizing reciprocity as extra information. The performance of proposed methods is comprehensively studied on eight realistic networks compared with several groups of benchmarks. Experimental results indicate that the proposed methods are more effective and robust than two state-of-the-art weighting methods and eight well-performing similarity indices.

INDEX TERMS Directed network, link prediction, reciprocal link, null model.

I. INTRODUCTION

Many complex systems can be naturally described with complex networks, where nodes represent individuals or entities while links represent their inner interactions [1]. Complex network is considered to be a powerful tool to explore fundamental properties of such complex systems. An elementary task of complex network is link prediction, the goal of which is to estimate the existence likelihood of unobserved or future interactions among nodes [2]. As a classic graph mining problem, link prediction has found a wide range of practical applications in many fields. For example, in online social networks link prediction can help users find their potential friends on social platforms such as Twitter and Sina Weibo, increasing their loyalty to the platform in turn [3]. In biological networks, link prediction methods can accelerate the process of discovering hidden structures of protein-protein interaction networks, saving both time and cost [4]. In information retrieval, link prediction in bipartite networks is regarded as a typical issue of recommender systems [5].

The past two decades have witnessed a rapid development on link prediction. Plenty of methods have been proposed to solve this challenging problem. Among those methods, similarity indices are popular ones for their effectiveness and simpleness [6]. Similarity indices are based on the assumption that if two endpoints share more common properties, a connection between them is more likely to exist. Representative similarity indices include Common Neighbor [7], Adamic-Adar [8], Resource Allocation [9], Jaccard [10], Salton [10], Sørensen [11], Leicht-Holme-Newman (LHN-I) [12], Hub Promoted Index (HPI) [13], Hub Depressed Index (HDI) [9], CAR [14], Local Path (LP) [9], just to name a few. In our previous works, we took into account the resource allocation of higher order paths as well as the potential information capacity of neighbors, and proposed two similarity indices: Extended Resource Allocation (ERA) index [15] and Potential Information Capacity (PIC) index [16].

In the real world, numerous of complex networks are directed, where link orientations have certain meanings.

The associate editor coordinating the review of this manuscript and approving it for publication was Chao-Yang Chen.

Twitter is a typical directed network which depicts the who-follow-whom relationships among users. Food webs are also directed networks presenting predation relationships in different species. For simplicity, most existing works on link prediction would assume the networks to be undirected. However, neglecting link orientations may lead to certain accuracy loss in link prediction, making predicting both the existence and orientation of links in directed networks an urgent and vital task in network science [17]. Schall et al. [18] proposed a statistical metric called Triadic Closeness (TC) to predict directed links. They use the probability that a given type of triad pattern will be closed as the likelihood of nonexistent links. Based on Schall's work, Bütün et al. [19] proposed a pattern-based supervised link prediction approach to enhance the accuracy of TC. Zhang et al. [20] took advantage of potential theory in physics to screen out most possible graph patterns in different directed networks. Combing potential theory with clustering and homophily mechanisms, they deduced a directed closure quad named "Bi-fan" (abbreviated as Bifan in the rest) as the most favored local structure in a majority of directed networks. They further designed Bifan index for link prediction. Wang et al. [21] extended the work of Zhou et al. [9] into directed networks and proposed a directed version of LP index. They added a ground node to the original network to take advantage of topological structures as much as possible. Lichtenwalter et al. [22] proposed an effective flow-based predicting algorithm for directed and weighted networks called PropFlow. It calculates the probability that a restricted random walk with l steps or fewer using link weights as transition probabilities. PropFlow is similar to rooted PageRank [10], but it is more localized of propagation and less sensitive to topological noise.

Relationships between nodes in directed networks can be classified into two categories: single directional link and reciprocal link. The latter one widely exists in all kinds of networks. For example, in information networks such as hyperlinks of websites, reciprocal links represent mutual linkage relationships between websites. In biological networks such as protein-protein interaction networks, a reciprocal link means that the connected proteins have mutual interactions with each other. Intuitively, reciprocal links have the ability to provide durable paths for information exchange inside the network. Thus they are assumed to be more informative than single directional ones [23]. Plenty of works have been done to explore the role of reciprocal links in directed networks. Garlaschelli and Loffredo [24] proposed a new measure of reciprocity along with a general framework to investigate the nonrandom presence of reciprocal link between two nodes. They found that networks of the same type always display similar values of reciprocity. Zhu et al. [25] studied the effect of reciprocal links on the function of real social networks, and found that reciprocal links play a more important role than single directional links in information diffusion process. Zhang et al. [26] focus on the impact of reciprocity on dynamical processes of networks. With the analysis on random walks in a scale-free directed and weighted network, they found evidence of the crucial role of reciprocity in random walks. Shang et al. [23] proposed a novel directional link prediction method to reveal the different roles of single directional links and reciprocal links on link formation. They concluded that two endpoints connected by reciprocal links are more likely to be linked to common neighbors than those connected by single directional links. Sett et al. [27] explored the effect of reciprocal links on triad formation in two large scale social networks: Facebook and Enron email network, and proposed three simple weighting mechanisms exploiting reciprocal links to enhance link prediction accuracy. Since most previous works mainly focus on social networks, we try to extend them and investigate the role of reciprocal links in different types of directed networks.

In social networks such as Twitter, if user A follows user B and continues to follow user C, a followee of user B, then the three users form a directed closure triad [28]. In some works this process is referred to as "link copying" [29]. The same structure are also found in other types of directed networks. Directed closure triads are regarded as fundamental blocks of directed networks and can be used for link prediction [30]. Lou et al. [28] studied how relationships develop in directed closure triads in social networks, and proposed a learning framework to formulate the problem of predicting directed closure triads into a graphical model. Brzozowski et al. [31] investigated WaterCooler network, an inner social network of HP Co., Ltd, and found that common neighbors forming "feed-forward-loop" triads are significant. Zhang et al. [32] extended undirected indices into directed versions based on the structure of "feed-forward-loop" and proposed Directed Common Neighbor (DCN) index, Directed Adamic-Adar (DAA) index, Directed Resource Allocation (DRA) index, and Directed Preferential Attachment (DPA) index.

Two interesting problems then arise: What role do reciprocal links exactly play in the formation of directed closure triads? And how can they be utilized to improve link prediction accuracy in directed networks? In this paper, we try to solve the above two problems in two ways. First we conduct a systematic investigation on the role of reciprocal links in four types of networks including social networks, information networks, infrastructure networks, and biological networks. Based on the results of an empirical test, we find solid evidence that reciprocal links are informative for link formation in directed closure triads. Then we propose two weighting mechanisms along with a set of reciprocal-aware weighted link prediction indices by utilizing reciprocity as extra information. The effectiveness and robustness of the proposed methods are validated and analyzed via comprehensive experiments on realistic networks.

The remainder of this paper is organized as follows. Preliminaries are presented in Section II. Section III introduces eight datasets for empirical test and validation. In Section IV, the role of reciprocal links is analyzed. Section V proposes two weighting mechanisms for link prediction. Section VI discusses experimental results. Section VIII draws conclusion of the paper.

II. PRELIMINARIES

A. PROBLEM DESCRIPTION

Considering a directed and unweighted network $D(V, E)$, denote V the set of nodes and E the set of links. For simplicity, reduplicate links and self-loops are deleted. The number of nodes and links in $D(V, E)$ are denoted as $|V|$ and $|E|$, respectively. $A = \{a_{ij}\}_{n\times n}$ represents the adjacency matrix, k_i^{in} denotes the in-degree of node i and k_i^{out} denotes its out-degree. Let $\Gamma_{\text{out}}(i)$ and $\Gamma_{\text{in}}(i)$ respectively be the set of out-going and in-coming neighbors. Denote $E^{\text{r}} = \{(u, v) \in E | (v, u) \in E\}$ as the set of reciprocal links. Notice that if $(u, v) \in E^{\text{r}}$, we call link (u, v) and (v, u) are both reciprocal links. The reciprocity coefficient $\rho = |E^{\text{r}}| / |E|$ is defined as the ratio of the number of reciprocal links to the total number of links [33]. Apparently, ρ is a real number in [0,1]. Let U be the universal set containing all possible links, the set of nonexistent links is $U - E$. For each nonexistent link $e(x, y) \in U - E, x, y \in V$, link prediction method assigns s_{xy} as a similarity score to quantify its existence likelihood. In directed networks, $s_{xy} \neq s_{yx}$. The unconnected nodes x, y are called seed nodes and link $e(x, y)$ is called candidate link.

B. SIMILARITY INDICES

Similarity indices are based on a fundamental assumption that "similarity breeds connection" [34]. Since node attributes are hard to get in practice, most similarity indices take advantage of the relative overlap between nodes' neighborhoods to predict missing links. Typically, the more "similar" two endpoints' neighborhoods are, the more likely they may establish a link. In most references, similarity indices are classified into two categorizes [2]: node-based indices and path-based indices. First we introduce four classic node-based similarity indices in directed networks which are widely used in practice for their effectiveness and simpleness.

1) Directed Common Neighbor (DCN) index [32]:
DCN index is an extension of CN index. It measures the number of common neighbors in the form of feed-forward-loops between two endpoints, denoted as:

$$s_{xy}^{\text{DCN}} = |\Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)| = \sum\nolimits_{z\in V} a_{xz} \cdot a_{zy} \quad (1)$$

2) Directed Adamic-Adar (DAA) index [32]:
DAA index is an extension of AA index which quantifies the features shared by two endpoints, and endows the rarer features with larger weights. It can characterize the neighborhood overlap between two endpoints, weighting the overlap of smaller neighborhoods more heavily.[1] It can be denoted as:

$$s_{xy}^{\text{DAA}} = \sum\nolimits_{z\in\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)} \frac{1}{\log(k_z^{\text{out}})} \quad (2)$$

3) Directed Resource Allocation (DRA) index [32]:
DRA index is an extension of RA index which considers the amount of given resources one endpoint has. It assumes that each node will distribute its resource equally among all out-going neighbors. The amount of received resource is considered to be relevant to the likelihood of a directed link. It can be denoted as:

$$s_{xy}^{\text{DRA}} = \sum\nolimits_{z\in\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)} \frac{1}{k_z^{\text{out}}} \quad (3)$$

4) Bifan index [20]:
Bifan index considers neighbors on paths of length 3. It assumes that if the existence of a directed link can lead to an increase of the number of Bifan motif, then this directed link is more likely to exist. The calculation of Bifan index can be denoted as:

$$\begin{aligned} s_{xy}^{\text{Bifan}} &= |\Gamma_{\text{in}}(\Gamma_{\text{out}}(x)) \cap \Gamma_{\text{in}}(y)| \\ &= \sum\nolimits_{z\in V} a_{xz} \cdot a_{z'z} \cdot a_{z'y} \end{aligned} \quad (4)$$

In addition, we introduce seven node-based similarity indices in directed networks, including Jaccard, Salton, Sørensen, LHN-I, HPI, HDI, and LP. Definitions of these indices are shown in Table 1.

TABLE 1. Definitions of state-of-art similarity indices for directed network.

Index	Definition	Reference
Jaccard	$\frac{\lvert\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)\rvert}{\lvert\Gamma_{\text{out}}(x)\cup\Gamma_{\text{in}}(y)\rvert}$	[10]
Salton	$\frac{\lvert\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)\rvert}{\sqrt{k_x^{\text{out}}\times k_y^{\text{in}}}}$	[10]
Sørensen	$\frac{2\lvert\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)\rvert}{k_x^{\text{out}}+k_y^{\text{in}}}$	[11]
LHN-I	$\frac{\lvert\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)\rvert}{k_x^{\text{out}}\times k_y^{\text{in}}}$	[12]
HPI	$\frac{\lvert\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)\rvert}{\min\{k_x^{\text{out}},k_y^{\text{in}}\}}$	[13]
HDI	$\frac{\lvert\Gamma_{\text{out}}(x)\cap\Gamma_{\text{in}}(y)\rvert}{\max\{k_x^{\text{out}},k_y^{\text{in}}\}}$	[2]
LP[1]	$\left(A^2\right)_{xy} + \alpha\left(A^3\right)_{xy}$	[9]

[1] α adjusts the weight of length-2 and length-3 paths.

Traditional node-based similarity indices calculate scores based on the binary adjacency matrix of the original network. Besides these indices, weighting methods transform the original adjacency matrix into a weighted one by utilizing certain structural or external properties [35]. In this case, the definition of common neighbors are changed, leading to extended forms of node-based indices such as DCN, DAA, and DRA. Local Naïve Bayes (LNB) model [36] is a well performing weighting method to improve the accuracy of basic similarity indices. It captures the roles of common neighbors and assign them with different weights. Reciprocal link count (RC) is another weighting method which counts the number of reciprocal links connecting the target node to their common neighbors [27]. The LNB and RC forms of DCN, DAA, DRA are presented in Table 2.

Different from node-based indices, path-based similarity indices conduct random walks on the network and take the arrival probability as the similarity score. PropFlow [22] is a popular path-based similarity index based on restricted random walks in no more than l steps. The restricted walk selects

[1] In [32], only out-going neighbors are utilized for calculation, while the effect of in-coming neighbors is considered insignificant.

links based on terminates when it reaches to another node or revisit any node, producing scores as the estimation of their existence likelihood. For two endpoints x and y, the score of PropFlow index PF(x, y) is [21]

$$\mathrm{PF}(x,y) = \mathrm{PF}(a,x)\frac{a_{xy}}{\sum_{k\in\Gamma_{\mathrm{out}}(x)} a_{xk}} \tag{5}$$

where k is x's neighbor whose depth is greater than x from the starting node. a is the previous node of x on a random walk path. When x is the starting node, PF(a, x) = 1.

C. EVALUATION METRICS

Link prediction in directed and unweighted networks can be regarded as a binary classification problem. To evaluate the accuracy of link prediction methods, the observed links in E are first randomly divided into two parts [37]: training set E^{T} and probe set E^{P}, as shown in Fig. 1. The training set can be regarded as the given information of link prediction methods,

TABLE 2. Definitions of RC and LNB forms of DCN, DAA and DRA.

Index	Definition
LNB-DCN[1]	$\lvert\Gamma_{\mathrm{out}}(x)\cap\Gamma_{\mathrm{in}}(y)\rvert\log d+\sum_{z\in\Gamma_{\mathrm{out}}(x)\cap\Gamma_{\mathrm{in}}(y)}\log R_z$
LNB-DAA	$\sum_{z\in\Gamma_{\mathrm{out}}(x)\cap\Gamma_{in}(y)}\frac{1}{\log k_z^{\mathrm{out}}}(\log d+\log R_z)$
LNB-DRA	$\sum_{z\in\Gamma_{\mathrm{out}}(x)\cap\Gamma_{in}(y)}\frac{1}{k_z^{\mathrm{out}}}(\log d+\log R_z)$
RC-DCN[2]	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} r(x,z)+r(z,y)$
RC-DAA[3]	$\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{r(x,z)+r(z,y)}{\log(1+s_z)}$
RC-DRA	$\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{r(x,z)+r(z,y)}{1+s_z}$

[1] $R_a=(N_{\Delta a}+1)/(N_{\wedge a}+1)$ is the role function, where $N_{\Delta a}$ and $N_{\wedge a}$ are respectively the number of connected and disconnected node pairs whose common neighbors include a. $d=M/M^{\mathrm{T}}-1$, where $M=|V|(|V|-1)/2$ and $M^{\mathrm{T}}=|E^{\mathrm{T}}|$. E^{T} is the training set.
[2] $r(x,y)=1$ when x and y are connected with single directional link, and $r(x,y)=2$ when x and y are connected with reciprocal link.
[3] $s_z=\sum_{z'\in\Gamma(z)} r(z,z')$ is the additive strength of node z.

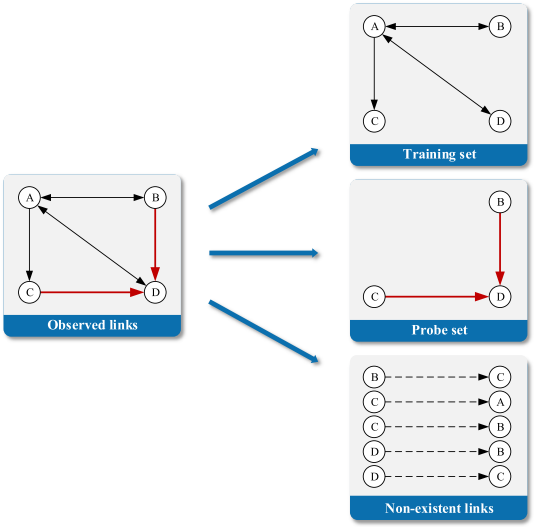

FIGURE 1. Diagram of dividing E into training set E^{T} and probe set E^{P} [37].

while the probe set is used for evaluating prediction accuracy. Let f be the partition ratio of training set, $E=E^{\mathrm{T}}\cup E^{\mathrm{P}}$, $E^{\mathrm{T}}\cap E^{\mathrm{P}}=\emptyset$, $|E^{\mathrm{T}}|=f\cdot|E|$ and $|E^{\mathrm{P}}|=(1-f)\cdot|E|$.

We choose two standard evaluation metrics to quantify the prediction accuracy: area under the receiver operating characteristic curve (AUC) [38] and precision [2]. AUC evaluates link prediction methods from an overall perspective while precision only concerns a few top ranked predictions.

AUC metric calculates the area under the receiver operating characteristic (ROC) curve. When the abscissa stands for the false positive rate and the ordinate stands for the true positive rate, a ROC curve can be drawn. Statistically, the area under ROC should be between 0.5 and 1. If AUC is greater than 0.5, we can suggest that a link prediction method is effective. If the area equals to 0.5, then the method is invalid. The case that the area is less than 0.5 is unrealistic [39]. A simplified way of estimating AUC in link prediction is to calculate the probability that the score of a randomly chosen missing link is higher than a randomly chosen nonexistent link. At each step, a missing link and a nonexistent link are selected randomly, and their similarity scores are compared. If among n independent comparisons, scores of missing links are higher for n' times and equal to those of nonexistent links for n'' times, AUC value is

$$\mathrm{AUC}=\frac{n'+0.5n''}{n} \tag{6}$$

Fig. 2 depicts a sample of calculating AUC. In this case, 4 nonexistent links (h, c, f, g) and 4 existent links (a, b, d, e) are chosen. The existent links among 16 pairs have higher scores than those of nonexistent links in 11 pairs: (a, h), (a, c), (a, f), (a, g), (b, c), (b, f), (b, g), (d, f), (d, g), (e, f), (e, g), and the score of link b is equal to a nonexistent link h. Therefore, AUC value is $(11+1\times0.5)/16\approx0.719$.

Different from AUC, precision metric focuses on the accuracy of top ranked predicted links. In practice, assuming among L predicted links, m of them can be found in the probe set, precision value is then calculated as

$$\mathrm{Precision}=m/L \tag{7}$$

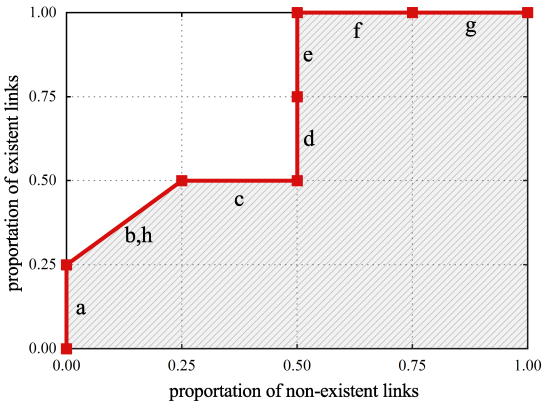

FIGURE 2. Diagram of calculating AUC.

TABLE 3. Basic statistics of twelve real-world directed networks.

No.	Network	Type	$\|V\|$	$\|E\|$	$\langle k\rangle$	ρ	C	$\langle d\rangle$	κ
1	ADO	social	2,539	12,969	10.22	38.76	14.20	5.30	0.251
2	HIG	social	70	366	10.46	50.28	40.40	3.51	0.083
3	PB	information	1,224	19,025	31.09	24.26	22.60	3.29	-0.221
4	EMA	information	2,029	39,264	38.70	71.12	8.90	3.98	-0.307
5	ATC	transport	1,226	2,615	4.27	15.70	6.39	8.05	-0.015
6	USA	transport	1,574	28,236	35.88	78.06	38.40	3.85	-0.113
7	CELE	biological	453	4,596	20.29	16.80	12.40	3.03	-0.226
8	FIG	biological	2,239	6,452	5.76	0.62	0.76	4.83	-0.331

Parameter L determines the number of links concerned. To compare among networks with different scales, we set L as proportional to the total number of links in each network.

III. DATASET

Eight directed networks from different fields in the real world are introduced for empirical test and validation. Only weakly connected component [40] of each network is concerned. A brief introduction of these realistic networks is described as follows:

1) Adolescent health (ADO) [41]: A friendship network created from a survey in 1994, where each participant was asked to list his/her five best female and five male friends. Nodes represent participants and links represent friendships.
2) High-school (HIG) [42]: A friendship network between boys in a high-school in Illinois. Each boy was asked to list his friends in 1957 and 1958. This dataset aggregates the results from both dates.
3) Political blogs (PB) [43]: A directed network of hyper-links among weblogs on US politics, recorded in 2005 by Adamic and Glance.
4) Email (EMA) [44]: An email communication network of employees in a European research institution. Nodes represent employees and links represent emails.
5) Air traffic control (ATC) [45]: A flight network constructed from the USA's Federal Aviation Administration National Flight Data Center (NFDC). Nodes represent airports or service centers and links represent preferred routes recommended by the NFDC.
6) US airports (USA) [46]: A directed network of flights between US airports in 2010. Each link represents a flight route from one airport to another.
7) Celegans (CELE) [47]: A metabolic network of the roundworm caenorhabditis elegans (C. elegans). Nodes represent metabolites (e.g., proteins) and links represent interactions between them.
8) Figeys (FIG) [48]: A network of interactions between proteins in Homo sapiens, from the first large-scale study of protein-protein interactions in Human cells using a mass spectrometry-based approach.

Table 3 presents the basic statistics of the introduced eight networks. $\langle k\rangle$ is the average node degree. ρ is the reciprocity coefficient [33]. C is the average clustering coefficient [49]. $\langle d\rangle$ is the 90-percentile effective diameter [50]. κ is the assortativity coefficient [51].

IV. EMPIRICAL TEST

In this section, an empirical test on realistic networks is designed to investigate the role of reciprocal links in the formation of directed closure triads. To address the importance of reciprocal links, we take advantage of a universal tool in network analyzing: null model.

In network science, null models are defined as a set of randomized graphs which preserve certain structural features of the original graph. Null model is often used as a term of comparison, to verify whether a graph displays some feature or not. In previous works, null models have been well exploited either to screen out significant network motifs in complex networks by comparing the count of subgraphs in the original graph and null models [52], or to discover community structures by comparing the modularity [53]. Here in order to reveal the role of reciprocal links in directed closure triads, we introduce a new type of null model for directed networks [27].

To facilitate our description, we first introduce the definition of underlying graph.

Definition 1: (Underlying graph) For a directed network $D(V, E)$, the underlying graph $U(V, E')$ of D is the undirected network created maintaining all nodes in V, and replacing all links in E with undirected links.

For a given directed network $D(V, E)$, we call $D_{\text{null}}(V, E_{\text{rand}})$ a reciprocal null model (RNM) of D, when $|E| = |E_{\text{rand}}|$ and $|E^{\text{r}}| = \left|E^{\text{r}}_{\text{rand}}\right|$, E^{r} and $E^{\text{r}}_{\text{rand}}$ are the set of reciprocal links in D and D_{null}, respectively. Apparently, in D_{null}, the directions of links are randomized while keeping the total number of directed links and reciprocal links. For a given directed network $D(V, E)$, its corresponding RNMs can be generated with the following steps:

1) Generate the underlying graph $U(V, E')$ of the given network $D(V, E)$. In this case, the reciprocity coefficient is $\rho = \left(|E| - \left|E'\right|\right)/\left|E'\right|$.
2) Randomly assign a one-way direction for each undirected link in $U(V, E')$. Let E'' be the new set of links.
3) Randomly select $|E| - \left|E'\right|$ links in E''. For each selected link, assign a reverse link between its endpoints to form a reciprocal link. The constructed RNM is denoted as $D_{\text{null}}(V, E_{\text{rand}})$, where E_{rand} is the set of reassigned directed links.

Fig. 3 presents the diagram of generating RNMs for a given directed network. In our empirical test, 1000 independent realizations of RNMs are generated. Then we are able to

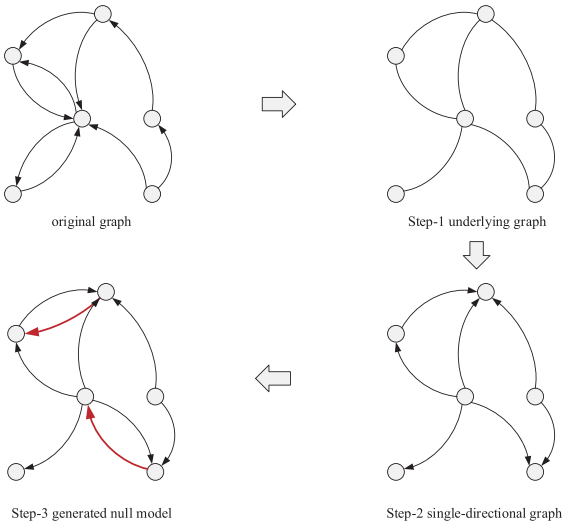

FIGURE 3. Diagram of generating RNMs for a given directed network.

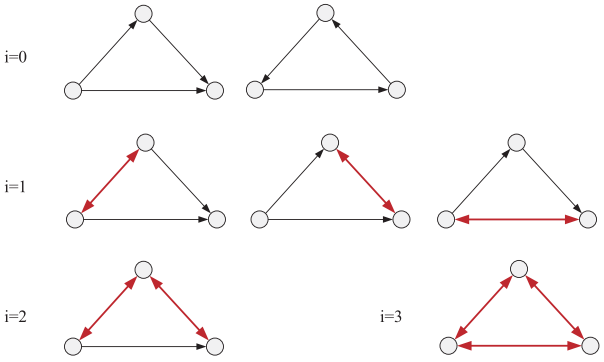

FIGURE 4. Diagram of possible directed closure triads with i reciprocal links (reciprocal links are marked red).

compare the number of different directed closure triads in the original network and RNMs. To address the role of reciprocal links, all possible directed closure triads are classified according to the number of reciprocal links they own, as shown in Fig. 4. For a directed closure triad, the possible number of reciprocal links is $\{0, 1, 2, 3\}$. Let N_i^D and N_i^{null} respectively be the number of directed closure triads with i reciprocal links in the given network D and its corresponding RNMs. The significance of subgraphs is measured by a standard metric called Z score [52], defined as

$$Z_i = \frac{N_i^D - \overline{N_i^{\text{null}}}}{\text{std}_i} \tag{8}$$

where $\overline{N_i^{\text{null}}}$ is the average number of directed closure triads with i reciprocal links under 1000 realizations of RNMs. std_i is the respective standard deviation.

Table 4 shows Z scores of directed closure triads with i reciprocal links in eight realistic networks introduced in Section III. From the result we can see that in all eight networks, Z score tends to be larger when i increases. It suggests that directed closure triads with a higher number of

TABLE 4. Z scores of directed closure triads with i ($i = 0, 1, 2, 3$) reciprocal links in eight networks.

Data	$\|E\|$	$\|E^{\text{r}}\|$	Z_0	Z_1	Z_2	Z_3
ADO	12,969	5,027	-17.890	-23.858	16.284	42.813
HIG	366	184	-2.309	-6.947	-0.417	8.987
PB	19,025	4,615	-25.450	-84.575	31.529	122.904
EMA	39,264	27,925	-15.564	-31.316	-14.218	50.367
ATC	2,615	411	-7.919	-2.173	14.099	10.917
USA	28,236	22,041	-12.504	-25.352	-85.077	51.536
CELE	4,596	772	-8.688	-11.875	6.879	8.366
FIG	6,452	40	-14.731	-8.924	22.429	220.012

reciprocal links are more significant in the original network compared with those in its corresponding RNMs. In highly reciprocal networks such as USA and EMA, even though Z_0 is bigger than Z_1 and Z_2, it is much smaller than Z_3. This is because these networks contain more reciprocal links. However, in networks with small reciprocity coefficient such as ATC, CELE, and FIG, directed closure triads with at least one reciprocal link are still more significant than those with no reciprocal links. For example, in FIG, whose reciprocal coefficient is 0.62, Z_3 is 220.012 while Z_0 is -14.731. This indicates that in these networks, directed closure triads with more reciprocal links are preferred.

To make the comparison more intuitive among networks with different scales, we further calculate the significance profile (SP) [54] value based on Z score. SP value is designed for analyzing superfamilies of unrelated networks by normalizing Z score into range [-1,1]. It emphasizes the relative significance of subgraphs instead of their absolute significance. SP value is defined as

$$\text{SP}_i = \frac{Z_i}{\sqrt{\sum_i Z_i^2}} \tag{9}$$

Fig. 5 shows the SP values of directed closure triads with i reciprocal links in eight realistic networks. From the results we can clearly observe the tendency that SP values increase with the value of i. When there are 0 or 1 reciprocal links in the triad, SP values are negative, which means the significance of these triads are less than completely pure chance. It indicates that the presence of directed closure triads with more than one reciprocal links is evident in different types of directed networks.

V. THE PROPOSED METHOD

The results in Table 4 and Fig. 5 in the empirical test motivate us that utilizing the reciprocate nature of directed closure triads may lead to improvement on link prediction accuracy. According to the structure of directed closure triads, two types of reciprocal links can be defined, namely indirect reciprocal link and direct reciprocal link [55].

As shown in Fig. 6, node x and y can either establish a direct reciprocal link, or establish two reciprocal links via their common neighbor z. Apparently the two types of reciprocal links have distinct capacities to transmit information. Based on the two types of reciprocal links, we propose an

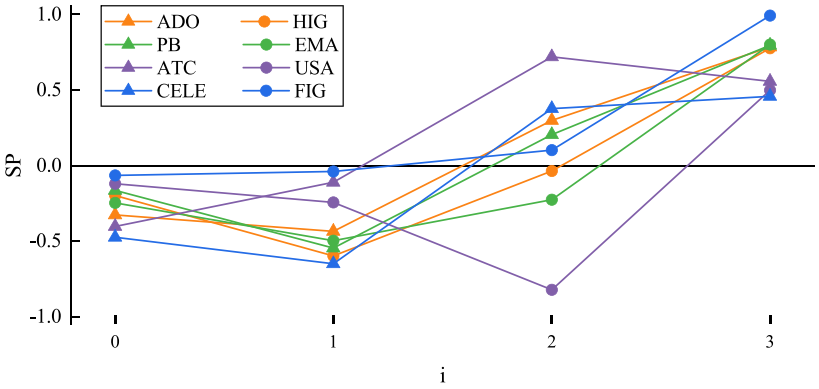

FIGURE 5. SP values of directed closure triads with $i(i = 0, 1, 2, 3)$ reciprocal links in different networks.

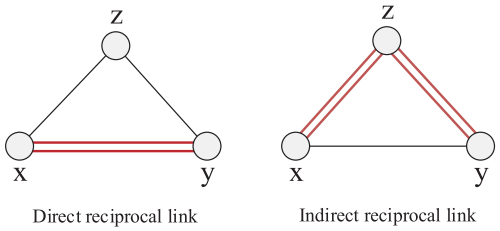

FIGURE 6. Diagram of indirect reciprocal link and direct reciprocal link.

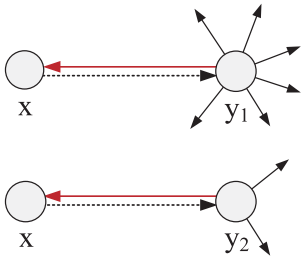

FIGURE 7. The effect of node degree on transmitted information via reciprocal link. Dotted line denotes possible link and colored line denotes reverse link.

indirect reciprocity-aware weighting method and a direct reciprocity-aware weighting method, respectively.

A. INDIRECT RECIPROCITY-AWARE WEIGHTING METHOD

The weight of links represents the strength of relationships between endpoints. For example, in social networks link weight depicts strength of social ties, while in biological networks link weight represents strength of interactions. Here we utilize the reciprocate nature in directed networks to calculate the strength of relationships. Intuitively, a reciprocal link represents a bidirectional information exchange between two nodes, and it can potentially provide more information than single directional links. Therefore, the strength of reciprocal links is considered to be stronger. A simple way to quantify the strength of reciprocal links is to add the effect of reverse link on the single directional link, as

$$w_{xy} = a_{xy} + \lambda \cdot a_{yx} \tag{10}$$

where λ is a tuning parameter adjusting the effect of reverse link. Since the effect of reverse link is related to multiple factors, λ is obviously not a constant.

The value of parameter λ reflects the amount of extra information an reverse link can provide. Typically, it is determined by both global and local properties of the entire network. On one hand, if there are more reciprocal links in the network, the existence of reverse link tends to be more significant. In this case, parameter λ is relevant to the reciprocity coefficient ρ. On the other hand, considering two seed nodes x and y, $e(x, y)$ is the candidate link, when $e(y, x) \in E$, the out-degree of end node y then determines the value of transmitted information through the reverse link $e(y, x)$. As shown in Fig. 7, when node y_1 has many out-going neighbors, the information transmitted to node x is less meaningful. On the contrary, when node y_2 has only a few out-going neighbors, reverse link $e(y, x)$ provides more valuable information to node x. Take Twitter as a simple example to illustrate this phenomenon, where users follow each other based on their interests. When user x has a follower y_1 who is an active user (the one who follows numerous of other users), x and y_1 are less likely to share common interests because y_1 could be one of the advertisers or artificial followers called "zombies". However, when x has a follower y_2 who only follows a few users including x, they have higher chance to be potential friends. The same phenomenon is also observed in other types of realistic networks [9], [15], [28].

Based on the discussions above, we define λ to be proportional to the reciprocity coefficient ρ and inverse proportional to the out-degree of target node y. The weighting mechanism of directed links is then denoted as

$$w_{xy} = a_{xy} + \rho \cdot \frac{a_{yx}}{k_y^{\text{out}}} \tag{11}$$

We can further modify the original adjacency matrix into a weighted one by using (11), where weights represent not only the connectivity but also the mutual information exchange between endpoints. In this case, the original problem turns into predicting missing links in weighted networks. Plenty of methods for link prediction in weighted networks have

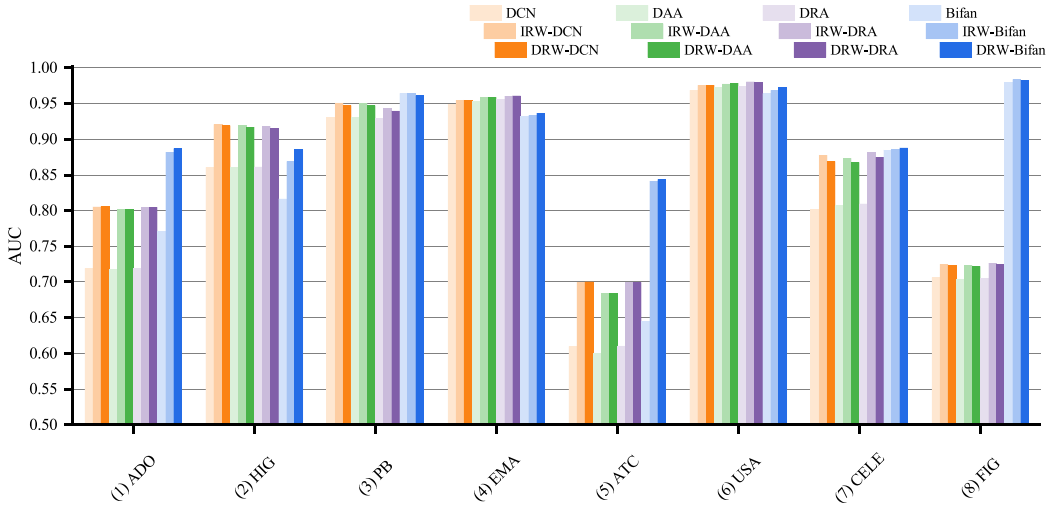

FIGURE 8. AUC values of proposed methods compared with basic similarity indices in eight networks.

been designed. Among them, the reliable-route weighting method shows its efficiency and effectiveness [56]. Based on the reliable-route weighting method, we construct a set of indirect reciprocity-aware weighting indices (IRW) as

1) IRW-DCN:

$$s_{xy}^{\mathrm{IRW-DCN}} = \sum\nolimits_{z\in V} w_{xz} \cdot w_{zy} \tag{12}$$

2) IRW-DAA:

$$s_{xy}^{\mathrm{IRW-DAA}} = \sum\nolimits_{z\in V} \frac{w_{xz} \cdot w_{zy}}{\log(1+s_z)} \tag{13}$$

where $s_z = \sum_{z'\in\Gamma(z)} w_{zz'}$ is the strength of z. We use $log(1 + s_z)$ instead of $log(s_z)$ to avoid negative values.

3) IRW-DRA:

$$s_{xy}^{\mathrm{IRW-DRA}} = \sum\nolimits_{z\in V} \frac{w_{xz} \cdot w_{zy}}{s_z} \tag{14}$$

where $s_z = \sum_{z'\in\Gamma(z)} w_{zz'}$ is the strength of node z.

4) IRW-Bifan:

$$s_{xy}^{\mathrm{IRW-Bifan}} = \sum\nolimits_{z\in V} w_{xz} \cdot w_{z'z} \cdot w_{z'y} \tag{15}$$

Notice that, according to the definition of reliable-route methods, link weights should be within range [0,1] in order to describe the probability that a link is safe for data transmission [56]. Therefore, before the calculation of similarity scores, link weights are normalized as

$$w' = f(w) \tag{16}$$

Reference [56] tested multiple normalization functions and found that weights normalized by logistic and exponential functions result in the highest precision values, for they can model inherent linkage likelihood of node pairs from the original weights. Here we choose the exponential function $f = e^{-\frac{1}{w}}$ for normalization.

B. DIRECT RECIPROCITY-AWARE WEIGHTING METHOD

Considering two endpoints x and y of a directed closure triad (x, z, y), when $e(y, x) \in E$, the possibility of the existence of $e(x, y)$ should be higher since reciprocal links are more informative to link prediction than single directional ones. Therefore, for each IRW index, we define its corresponding direct reciprocal-aware weighting (DRW) counterpart as:

$$s_{xy}^{\mathrm{DRW}} = s_{xy}^{\mathrm{IRW}} + \lambda' \cdot s_{yx}^{\mathrm{IRW}} \tag{17}$$

where s_{xy}^{IRW} is similarity score of IRW index, λ' adjusts the effect of reverse link. Similar to λ in (11), we define λ' as proportional to the reciprocity coefficient and inverse proportional to the out-degree of target node. The DRW-based indices are then denoted as:

$$s_{xy}^{\mathrm{DRW}} = s_{xy}^{\mathrm{IRW}} + \rho \cdot \frac{s_{yx}^{\mathrm{IRW}}}{k_y^{\mathrm{out}}} \tag{18}$$

where ρ is the reciprocity coefficient.

VI. RESULTS AND DISCUSSIONS

A. ACCURACY OF PROPOSED METHODS

The performance of proposed methods in comparison with their basic counterparts is analyzed first. Fig. 8 and Fig. 9 respectively show the AUC and precision values of basic, IRW-based, and DRW-based similarity indices in eight networks. The partition ratio of training set is $f = 0.9$, and parameter L in the calculation of precision value in (7) is set to 1% of $|E|$. Each value is generated by averaging the results of 30 independent implementations for each network.

In general, an obvious improvement on both AUC and precision can be observed. Especially in ADO and ATC, AUC values of proposed methods are averagely higher than basic ones by 8%. In HIG, precision values of DRW-based indices are approximately twice of basic ones. Even in networks with small reciprocity coefficients such as CELE and FIG, both the AUC and precision values of proposed methods

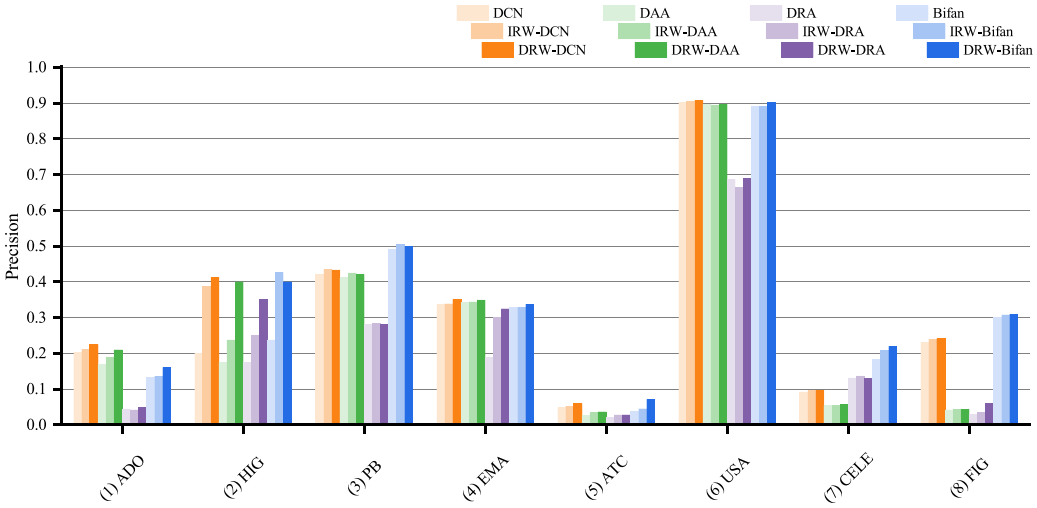

FIGURE 9. Precision values of proposed methods compared with basic similarity indices in eight networks.

TABLE 5. AUC and precision values of DCN series. The left part presents AUC values and the right part presents precision values. Index with the best performance in each network is marked in bold.

Network	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG
DCN	0.715	0.869	0.929	0.948	0.610	0.969	0.797	0.704	0.189	0.158	0.419	0.339	0.059	0.903	0.104	0.230
RC-DCN	0.715	0.872	0.928	0.948	0.610	0.969	0.797	0.704	0.196	0.175	0.427	0.337	0.066	0.903	0.103	0.227
LNB-DCN	0.716	0.866	0.921	0.953	0.611	0.970	0.774	0.506	0.184	0.154	**0.455**	0.339	0.028	0.904	0.090	0.052
LRW-DCN	0.805	**0.921**	0.948	0.953	0.705	0.975	**0.857**	0.708	0.194	0.177	0.425	0.342	**0.075**	0.904	0.104	0.214
DRW-DCN	**0.805**	0.918	**0.948**	**0.954**	**0.705**	**0.975**	0.855	**0.709**	**0.200**	**0.185**	0.428	**0.347**	0.062	**0.906**	**0.105**	**0.246**

TABLE 6. AUC and precision values of DAA series. The left part presents AUC values and the right part presents precision values. Index with the best performance in each network is marked in bold.

Index	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG
DAA	0.713	0.870	0.929	0.952	0.602	0.972	0.802	0.704	0.169	0.146	0.408	0.344	0.030	0.896	0.050	0.039
RC-DAA	0.714	0.873	0.928	0.951	0.602	0.971	0.801	0.704	0.158	**0.195**	0.411	0.317	0.045	0.888	0.056	0.038
LNB-DAA	0.714	0.864	0.922	0.955	0.603	0.972	0.775	0.506	0.160	0.150	0.436	0.364	0.036	0.891	0.059	0.035
LRW-DAA	0.798	0.917	**0.950**	0.957	0.690	0.977	**0.872**	**0.724**	0.174	0.159	0.412	0.347	0.048	0.894	0.077	**0.041**
DRW-DAA	**0.798**	**0.917**	0.948	**0.958**	**0.690**	**0.978**	0.864	0.703	**0.190**	0.182	**0.415**	**0.350**	**0.050**	**0.898**	**0.078**	0.035

TABLE 7. AUC and precision values of DRA series. The left part presents AUC values and the right part presents precision values. Index with the best performance in each network is marked in bold.

Index	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG
DRA	0.716	0.869	0.929	0.955	0.610	0.974	0.804	0.704	0.037	0.145	0.274	0.190	0.038	0.691	0.101	0.023
RC-DRA	0.715	0.874	0.930	0.954	0.609	0.973	0.804	0.704	**0.147**	0.162	0.253	0.337	0.031	0.702	0.096	0.023
LNB-DRA	0.716	0.868	0.922	0.954	0.611	0.974	0.774	0.509	0.041	0.162	0.186	0.121	0.024	0.600	0.092	0.020
LRW-DRA	**0.802**	0.916	0.946	0.959	**0.704**	0.979	**0.880**	**0.726**	0.030	0.158	0.269	0.279	**0.034**	0.659	0.102	0.026
DRW-DRA	0.800	**0.916**	**0.946**	**0.960**	0.702	**0.981**	0.873	0.703	0.040	**0.174**	**0.273**	**0.369**	0.033	**0.717**	**0.102**	**0.026**

are significantly improved. It implies that in these networks, reciprocal links play an important role in providing extra information and stimulate link formation.

Subsequently, we compare the proposed methods with two aforementioned weighting methods in Section II-B: RC and LNB. Performance comparison of basic, RC-based, LNB-based, IRW-based, and DRW-based similarity indices is presented in Table 5, 6, and 7, respectively. In most networks, weighting methods are able to improve the accuracy of basic indices at different levels. For example, the precision value of LNB-DCN is 3.6% higher than that of DCN in PB. In HIG, the precision value of RC-DAA is 4.9% higher than DAA. Nevertheless, in most networks, both RC-based and LNB-based indices perform nearly the same. The proposed methods, however, are able to achieve more accurate predictions in all eight networks. In some networks, the improvement is quite obvious. For example, in ADO, the AUC values of LRW-DCN and DRW-DCN are higher than DCN by approximately 9.0%. In most networks, the proposed methods get the best performance compared with basic, RC-based, and LNB-based counterparts.

Table 8 shows the performance comparison of proposed methods and eight state-of-the-art similarity indices including those listed in Table 1 and PropFlow. In general, Jaccard,

TABLE 8. Performance comparison of proposed methods and other state-of-art similarity indices. The left part presents AUC values and the right part presents precision values. Index with the best performance in each network is marked in bold.

Index	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG	ADO	HIG	PB	EMA	ATC	USA	CELE	FIG
Jaccard	0.715	0.874	0.909	0.940	0.609	0.950	0.795	0.710	0.077	0.178	0.002	0.244	0.000	0.001	0.008	0.015
Salton	0.715	0.874	0.908	0.943	0.609	0.952	0.800	0.710	0.055	0.180	0.002	0.174	0.000	0.001	0.012	0.015
Sørensen	0.715	0.872	0.908	0.941	0.608	0.951	0.795	0.710	0.077	0.178	0.002	0.244	0.000	0.001	0.008	0.015
LHN-I	0.714	0.866	0.880	0.889	0.608	0.872	0.789	0.709	0.027	0.088	0.002	0.001	0.000	0.000	0.006	0.015
HPI	0.714	0.871	0.902	0.926	0.609	0.924	0.803	0.708	0.024	0.140	0.017	0.006	0.016	0.013	0.140	0.048
HDI	0.714	0.872	0.907	0.936	0.607	0.948	0.794	0.710	0.124	0.168	0.002	0.210	0.000	0.001	0.010	0.017
LP[1]	0.781	0.884	0.962	0.952	0.701	0.976	0.864	0.804	0.172	0.180	0.414	0.334	0.038	0.896	0.098	0.202
PropFlow[1]	0.868	0.879	0.938	0.916	0.840	0.937	0.869	0.961	0.030	0.144	0.016	0.033	0.008	0.100	0.204	0.034
IRW-DCN	0.805	**0.921**	0.948	0.953	0.705	0.975	0.857	0.708	0.194	0.177	0.425	0.342	0.075	0.904	0.104	0.214
DRW-DCN	0.805	0.918	0.948	0.954	0.705	0.975	0.855	0.709	**0.200**	**0.185**	0.428	0.347	0.062	**0.906**	0.105	0.246
IRW-DAA	0.798	0.917	0.950	0.957	0.690	0.977	0.872	0.724	0.174	0.159	0.412	0.347	0.048	0.894	0.077	0.041
DRW-DAA	0.798	0.917	0.948	0.958	0.690	0.978	0.864	0.703	0.190	0.182	0.415	0.350	0.050	0.898	0.078	0.035
IRW-DRA	0.802	0.916	0.946	0.959	0.704	0.979	0.880	0.726	0.030	0.158	0.269	0.279	0.034	0.659	0.102	0.026
DRW-DRA	0.800	0.916	0.946	**0.960**	0.702	**0.981**	0.873	0.703	0.040	0.174	0.273	**0.369**	0.033	0.717	0.102	0.026
IRW-Bifan	0.878	0.875	0.963	0.933	0.843	0.967	0.887	0.984	0.126	0.139	0.500	0.330	0.076	0.891	**0.184**	0.319
DRW-Bifan	**0.883**	0.888	**0.963**	0.936	**0.845**	0.970	**0.888**	**0.984**	0.139	0.166	**0.502**	0.340	**0.079**	0.897	0.176	**0.319**

[1] Parameter $\alpha = 0.001$ in LP. Parameter $l = 5$ in PropFlow.

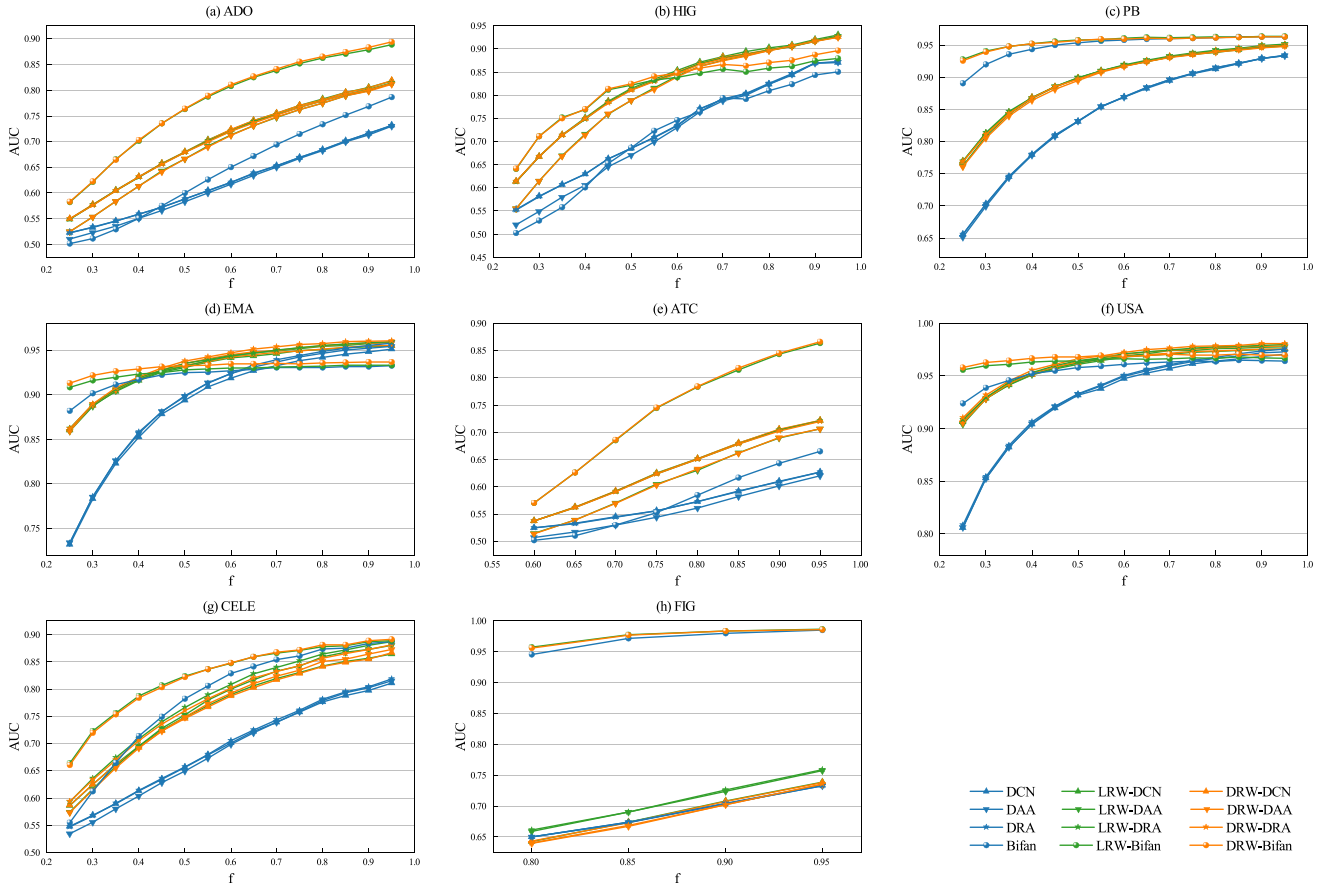

FIGURE 10. AUC values of proposed methods in eight networks with different sizes of training sets.

Salton, Sørensen, LHN-I, HPI, and HDI show almost the same accuracy in each network, because they all exploit common neighbors between two endpoints with distinct assumptions on neighbors' contributions. As quasi-local indices, LP and PropFlow utilizes more information from longer paths besides common neighbors, leading to obvious improvement on accuracy. Nevertheless, the proposed methods achieve better performance in most networks. DRW-Bifan has the highest AUC value in 5 out of 8 networks, and performs the best in PB, ATC, and FIG under precision metric. DRW-DRA and DRW-DCN also outperforms most benchmarks in other networks under both AUC and precision metrics.

Overall, the results suggest that taking into account the effect of reciprocal links can efficiently improve prediction

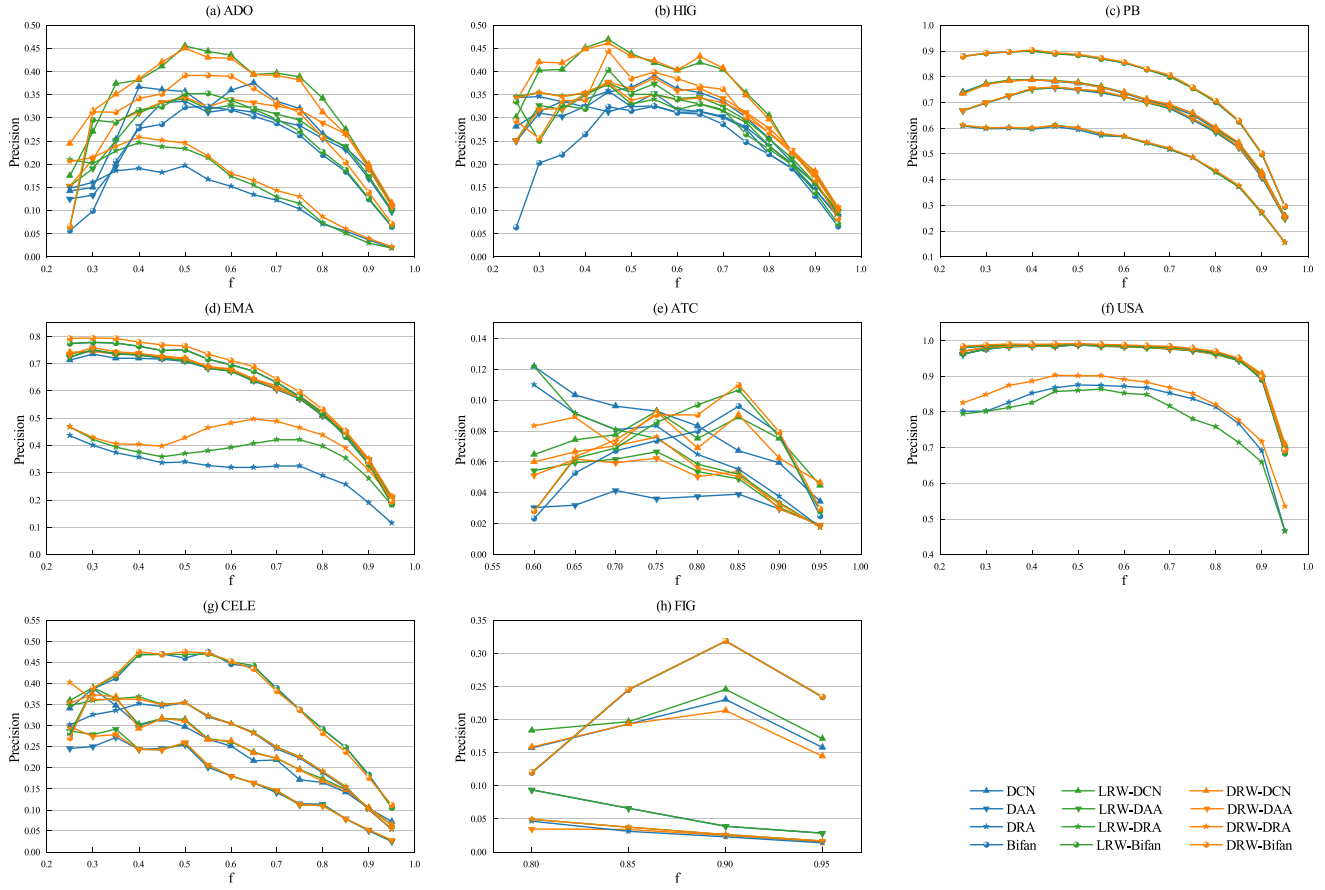

FIGURE 11. Precision values of proposed methods in eight networks with different sizes of training sets.

accuracy of four basic similarity indices on both AUC and precision. It provides more evidence for the results drawn from the empirical test in Section IV. Moreover, as a complement of Sett's conclusions [27], the results imply that the role of reciprocal links in different types of directed networks is both significant and informative, and reciprocal links can be well utilized to improve link prediction accuracy by using the weighting methods we propose.

B. ROBUSTNESS ANALYSIS ON THE SIZE OF TRAINING SET

To analyze the robustness of the proposed methods, we compare the performance with changes on the size of training set. The AUC and precision values of different methods when partition ratio of training sets f changes from 0.25 to 0.95 are respectively shown in Fig. 10 and Fig. 11. Each value is the average of 30 independent implementations with random divisions of training set and probe set. Parameter L in the calculation of precision value in (7) is set to 1% of $|E|$.

In Fig. 10, it is clear to see that in most networks, DRW-based and IRW-based methods have higher AUC values than basic methods under all partition ratios. In ADO, HIG, ATC, and CELE, the gaps between the proposed methods and basic ones are obvious. In PB, EMA, and USA, the AUC curves of DRW-based and IRW-based indices decrease more mildly than those of basic ones, indicating that the proposed methods are more robust against the size of training set in these networks. We also notice that in USA, basic indices such as DCN, DAA, and DRA only achieve 0.81 in AUC value when only 25% of links in the original network are observed. However, the corresponding DRW-based and IRW-based counterparts still get AUC values of approximately 0.91. We infer that reciprocal links are able to bring extra information to link prediction even when the knowledge from the partially observed network is limited.

In Fig. 11, improvements on robustness of precision values are also observed in some networks such as ADO, HIG, and USA. In HIG, when $f = 0.25$, the precision value of Bifan is 0.06. However, the precision values of DRW-Bifan and IRW-Bifan are increased by 28% and 23%. Nevertheless, we also notice that in PB and CELE, the precision values of proposed methods are nearly the same as basic ones.

VII. DATA AND CODE AVAILABILITY

The original data of eight realistic networks are available from: http://konect.uni-koblenz.de/networks/.

The source codes of the proposed methods and benchmarks are available from: https://github.com/Lee3Paul/LP_with_reciprocal_links.

VIII. CONCLUSION

Recent years have witnessed a rapid development of network science and link prediction. Predicting missing links in directed networks is considered to be both promising and challenging. In this paper, we investigate the role of reciprocal links in the formation of directed closure triads in different types of directed networks via an empirical test. Two weighting mechanisms along with eight weighted indices are then proposed based on four state-of-the-art similarity indices: DCN, DAA, DRA, and Bifan, by differentiating the effects of reciprocal link and single directional link. The proposed weighting mechanisms consider indirect and direct reciprocity between two endpoints in a directed closure triad. Both global property (i.e., the reciprocity coefficient) and local property (i.e., the out-degree of target node) are utilized to quantify the effect of reverse links. Experimental results on eight realistic networks from different fields indicate that the proposed methods outperform their counterparts under AUC and precision metrics. In addition, the proposed methods show better robustness on the size of training set.

In this work, we focus on predicting missing links in directed and unweighted networks. This work may be extended to weighted, multi-relational networks considering the influence of external information such as node attributes, prior knowledge, etc. Moreover, the role of reciprocal links in temporal directed networks may also be investigated, which can potentially provide new insights toward link prediction.

REFERENCES

[1] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, Jan. 2002.
[2] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Stat. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011.
[3] S.-Y. Teng, L.-P.-Y. Ting, M.-Y. Yeh, and K.-T. Chuang, "Worship prediction: Identify followers in celebrity-dived networks," *World Wide Web*, vol. 22, no. 1, pp. 347–373, Jan. 2019.
[4] C. Lei and J. Ruan, "A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity," *Bioinformatics*, vol. 29, no. 3, pp. 355–364, Feb. 2013.
[5] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems," *Phys. Rep.*, vol. 519, no. 1, pp. 1–49, Oct. 2012.
[6] S. Liu, X. Ji, C. Liu, and Y. Bai, "Similarity indices based on link weight assignment for link prediction of unweighted complex networks," *Int. J. Mod. Phys. B*, vol. 31, no. 02, Jan. 2017, Art. no. 1650254.
[7] G. Kossinets, "Effects of missing data in social networks," *Social Netw.*, vol. 28, no. 3, pp. 247–268, Jul. 2006.
[8] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Netw.*, vol. 25, no. 3, pp. 211–230, Jul. 2003.
[9] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, Oct. 2009.
[10] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci.*, vol. 58, no. 7, pp. 1019–1031, May 2007.
[11] T. Sørensen, "A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Biol. Skr.*, vol. 5, pp. 1–34, Jan. 1948.
[12] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 73, no. 2, Feb. 2006, Art. no. 026120.
[13] E. Ravasz, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002.
[14] 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," *Sci. Rep.*, vol. 3, no. 1, pp. 1–4, Dec. 2013.
[15] S. Liu, X. Ji, C. Liu, and Y. Bai, "Extended resource allocation index for link prediction of complex network," *Phys. A, Stat. Mech. Appl.*, vol. 479, pp. 174–183, Aug. 2017.
[16] X. Li, S. Liu, H. Chen, and K. Wang, "A potential information capacity index for link prediction of complex networks based on the cannikin law," *Entropy*, vol. 21, no. 9, p. 863, Sep. 2019.
[17] E. Bütün, M. Kaya, and R. Alhajj, "Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks," *Inf. Sci.*, vols. 463–464, pp. 152–165, Oct. 2018.
[18] D. Schall, "Link prediction in directed social networks," *Soc. Netw. Anal. Min.*, vol. 4, no. 1, p. 157, Dec. 2014.
[19] E. Bütün and M. Kaya, "A pattern based supervised link prediction in directed complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 525, pp. 1136–1145, Jul. 2019.
[20] Q.-M. Zhang, L. Lü, W.-Q. Wang, T. Zhou, "Potential theory for directed networks," *PLoS ONE*, vol. 8, no. 2, Feb. 2013, Art. no. e55437.
[21] X. Wang, X. Zhang, C. Zhao, Z. Xie, S. Zhang, and D. Yi, "Predicting link directions using local directed path," *Phys. A, Stat. Mech. Appl.*, vol. 419, pp. 260–267, Feb. 2015.
[22] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, 2010, p. 243.
[23] K.-K. Shang, M. Small, and W.-S. Yan, "Link direction for link prediction," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 767–776, Mar. 2017.
[24] D. Garlaschelli and M. I. Loffredo, "Patterns of link reciprocity in directed networks," *Phys. Rev. Lett.*, vol. 93, no. 26, Dec. 2004, Art. no. 268701.
[25] Y.-X. Zhu, X.-G. Zhang, G.-Q. Sun, M. Tang, T. Zhou, and Z.-K. Zhang, "Influence of reciprocal links in social networks," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e103007.
[26] Z. Zhang, H. Li, and Y. Sheng, "Effects of reciprocity on random walks in weighted networks," *Sci. Rep.*, vol. 4, no. 1, p. 7460, May 2015.
[27] N. Sett, S. R. Singh, and S. Nandi, "Exploiting reciprocity toward link prediction," *Knowl. Inf. Syst.*, vol. 55, no. 1, pp. 1–13, Apr. 2018.
[28] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding, "Learning to predict reciprocity and triadic closure in social networks," *ACM Trans. Knowl. Discov. Data*, vol. 7, no. 2, pp. 1–25, Jul. 2013.
[29] D. M. Romero and J. Kleinberg, "The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter," 2010, *arXiv:1003.2469*. [Online]. Available: https://arxiv.org/abs/1003.2469
[30] E. Bütün, M. Kaya, and R. Alhajj, "A new topological metric for link prediction in directed, weighted and temporal networks," in *Proc. ASONAM*, San Francisco, CA, USA, 2016, pp. 954–959.
[31] M. J. Brzozowski and D. M. Romero, "Who should I follow? Recommending people in directed social networks," in *Proc. Comput. Support. Coop. Work*, Jul. 2011, pp. 1–10.
[32] X. Zhang, C. Zhao, X. Wang, and D. Yi, "Identifying missing and spurious interactions in directed networks," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 9, Sep. 2015, Art. no. 507386.
[33] M. E. J. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 66, no. 3, pp. 17–20, 2002.
[34] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proc. WWW*. New York, NY, USA: ACM, 2008, pp. 915–924.
[35] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Comput. Surv.*, vol. 49, no. 4, pp. 1–33, Dec. 2016.
[36] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: A local naive Bayes model," *Europhysics Lett.*, vol. 96, no. 4, p. 48007, Nov. 2011.
[37] K.-K. Shang, M. Small, and W.-S. Yan, "Fitness networks for real world systems via modified preferential attachment," *Phys. A, Stat. Mech. Appl.*, vol. 474, pp. 49–60, May 2017.
[38] J. M. Hofman, A. Sharma, and D. J. Watts, "Prediction and explanation in social systems," *Science*, vol. 355, no. 6324, pp. 486–488, Feb. 2017.
[39] K.-K. Shang, T.-C. Li, M. Small, D. Burton, and Y. Wang, "Link prediction for tree-like networks," *Chaos*, vol. 29, no. 6, Jun. 2019, Art. no. 061103.
[40] G. Bianconi, N. Gulbahce, and A. E. Motter, "Local structure of directed networks," *Phys. Rev. Lett.*, vol. 100, no. 11, Mar. 2008.
[41] P. Massa, M. Salvetti, and D. Tomasoni, "Bowling alone and trust decline in social network sites," in *Proc. DASC*, Chengdu, China, 2009, pp. 658–663.

[42] J. S. Coleman, *Introduction to Mathematical Sociology*. London, U.K.: London Free Press, 1964.
[43] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: Divided they blog," in *Proc. LinkKDD*, New York, NY, USA, 2005, pp. 36–43.
[44] J. Leskovec and A. Krevl. *SNAP Datasets: Standford Large Network Dataset Collection*. Accessed: Jan. 2014. [Online]. Available: http://snap.standford.edu/data
[45] Federal Aviation Administration. *Air Traffic Control System Command Center*. Accessed: Apr. 2017. [Online]. Available: http://www.fly.faa.gov/
[46] J. Kunegis. *US Airports Network Dataset–KONECT*. Accessed: Sep. 2016. [Online]. Available: http://konect.uni-koblenz.de/networks/opsahl-usairport
[47] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode Caenorhabditis elegans," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 314, no. 1165, pp. 1–340, Nov. 1986.
[48] J. Leskovec and A. Krevl. *Human Protein (Figeys) Network Dataset–KONECT*. Accessed: Sep. 2016. [Online]. Available: http://konect.uni-koblenz.de/networks/maayan-figeys
[49] G. Fagiolo, "Clustering in complex directed networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 2, Aug. 2007, Art. no. 026107.
[50] C. R. Palmer, P. B. Gibbons, and C. Faloutsos, "ANF: A fast and scalable tool for data mining in massive graphs," in *Proc. KDD*. Edmonton, AB, Canada: ACM Press, 2002, p. 81.
[51] Q. Guo, T. Zhou, J.-G. Liu, W.-J. Bai, B.-H. Wang, and M. Zhao, "Growing scale-free small-world networks with tunable assortative coefficient," *Phys. A, Stat. Mech. Appl.*, vol. 371, no. 2, pp. 814–822, Nov. 2006.
[52] R. Milo, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.
[53] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113.
[54] R. Milo, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, Mar. 2004.
[55] K.-K. Shang, M. Small, X.-K. Xu, and W.-S. Yan, "The role of direct links for link prediction in evolving networks," *Europhysics Lett.*, vol. 117, no. 2, Jan. 2017, Art. no. 28002.
[56] J. Zhao, L. Miao, J. Yang, H. Fang, Q.-M. Zhang, M. Nie, P. Holme, and T. Zhou, "Prediction of links and weights in networks by reliable routes," *Sci. Rep.*, vol. 5, p. 12261, Jul. 2015.

JINSONG LI received the B.E. degree in information engineering from the PLA Strategic Support Force Information Engineering University (PLAIEU), Zhengzhou, Henan, China, in 2014, and the M.E. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2017. He is currently pursuing the Ph.D. degree with the National Digital Switching System Engineering and Technological Research Center (NDSC), Zhengzhou. His research interests include network science, social network analysis, data mining, indoor localization, and cyber security.

JIANHUA PENG received the M.E. degree in computer application, in 1995. He is currently a Professor, a Doctoral Supervisor, and also the Deputy Chief Engineer with the National Digital Switching System Engineering and Technological Research Center (NDSC). His research interests include network science, social network analysis, network switching, cyber security, and telecommunications.

SHUXIN LIU received the B.E. degree in communication engineering from the PLA Strategic Support Force Information Engineering University (PLAIEU), in 2009, and the M.S. and Ph.D. degrees from the National Digital Switching System Engineering and Technological Research Center (NDSC), in 2012 and 2016, respectively, where he is currently a Research Assistant. He is the author of more than 20 articles and more than ten inventions. His research interests include network evolution, link prediction, social network analysis, and communication network security.

XINSHENG JI received the B.E. degree in digital communication and the M.S. degree from Fudan University, Shanghai, China, in 1984, and the M.S. degree from the PLA Strategic Support Force Information Engineering University (PLAIEU), in 1991. He is currently pursuing the Eng.D. degree with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He is currently the Chief Engineer of the National Digital Switching System Engineering and Technological Research Center (NDSC). His major research interests include wireless communication, network security, network science, and signal processing. He has been a member of the Network and Communication Specialist Group for the China 863 High Technology Program. He has also been a Senior Member of the China Institute of Communication. He received the Outstanding Expert of State Award, in 2015.

XING LI received the B.E. and M.E. degrees in information engineering from the PLA Strategic Support Force Information Engineering University (PLAIEU), where he is currently pursuing the Ph.D. degree. He is currently an Assistant Researcher with the National Digital Switching System Engineering and Technological Research Center (NDSC). His research interests include link prediction, network science, and social network analysis.

XINXIN HU received the B.E. degree from the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, in 2017. He is currently pursuing the M.S. degree with the National Digital Switching System Engineering and Technological Research Center (NDSC). His major research interests include network science, next generation mobile network, and network robustness analysis.