Research Article

Identifying Missing and Spurious Interactions in Directed Networks

Xue Zhang, 1 Chengli Zhao, 1 Xiaojie Wang, 1 and Dongyun Yi 1,2

1 School of Science, National University of Defense Technology, Changsha 410073, China
2 State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China

Correspondence should be addressed to Chengli Zhao; chenglizhao@gmail.com

Received 31 August 2014; Accepted 24 October 2014

Academic Editor: Xiuzhen Cheng

Copyright © 2015 Xue Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recent years, the studies of link prediction have been overwhelmingly emphasizing on undirected networks. Compared with it, how to identify missing and spurious interactions in directed networks has received less attention and still is not well understood. In this paper, we make use of classical link prediction indices for undirected networks, adapt them to directed version which could predict both the existence and direction of an arc between two nodes, and investigate their prediction ability on six real-world directed networks. Experimental results demonstrate that those modified indices perform quite well in directed networks. Compared with bifan predictor, some of them can provide more accurate predictions.

1. Introduction

Network is an effective and efficient tool to describe real-world complex systems [1, 2], such as social, biological, traffic, and information systems, where nodes represent individuals, proteins, airports, web pages, and so forth and links denote the relations and interactions between them. The study of complex networks is an important growing field that attracts lots of attention from different branches of science. While making great efforts to understand the structural features and evolutionary mechanism of networks, scientists gradually realize that the inaccuracy and incompleteness of data sets is a significant obstacle to the research [3, 4]. To address this issue, link prediction algorithms have been adopted to extract the missing information, identify spurious interactions, and reconstruct network.

The problem of link prediction aims at estimating the likelihood of the existence of a link between two nodes in a given network based on the observed links [5], and it has a wide range of applications in the real-world. For example, in online social networks, very likely but not-yet-existent links can be recommended as promising friendship, which can help users in finding new friends; in biological networks, compared to blindly checking all possible protein-protein interactions, accurate prediction of the most likely existent ones can dramatically reduce the experimental cost; in e-commerce, with the help of recommendation systems, sellers enhance their sales by watching customers’ purchases and recommending other goods to them in which they may be interested [6]; in security domain, link prediction methods could be used to assist identifying groups of terrorist or criminals [7].

Since the link prediction problem is relevant for various domains, lots of algorithms have been proposed to solve it. The most widely used link prediction indices are the local similarity measures [8–12], for example, Common Neighbors, Jaccard index, Adamic-Adar index, and Resource Allocation index, which only require consideration of the local structure of the networks. They are favored for low complexity, low time consumption, and relatively high accuracy especially when the network is highly clustered [13].

Algorithm based on statistical inference is another branch in the study of link prediction, in which some organizing principles of the network structure, like hierarchical organizations or community structures, are often presupposed [7, 14, 15]. Using Bayes theorem, people can infer the underlying structure from observed network and take advantage of the knowledge to reliably identify both missing and spurious
interactions. Usually, the performance of this type of method is more accurate and robust, but it has an obvious drawback—the high computational complexity makes it infeasible for large-scale networks.

All the algorithms mentioned above are designed for undirected networks. However, in the real-world asymmetric interactions are widespread such as worldwide web, food web, neural network, email network, and citation network. Unfortunately, there has been little research focusing on how to identify missing and spurious interactions in directed networks. Until very recently, a new mechanism for the local organization of directed networks, potential theory, has been proposed [16]. Combining the potential theory with the clustering and homophily mechanisms, bifan structure which consists of 4 nodes and 4 directed links is deduced to be the most favored local motif structure and could be used directly as a well-performed missing link predictor in directed networks.

In this paper, we focus on the link prediction problem in directed networks. Our contribution is twofold: (1) Instead of proposing a brand-new directed link predictor, we make the best of classical prediction indices of undirected networks. We extend the most representative measures to directed networks. Our contribution is twofold: (1) Instead of proposing a brand-new directed link predictor, we make the best of classical prediction indices of undirected networks. To investigate the performance of these modified indices, we design simulation experiments on five real biological and technological directed networks, the results of which vividly demonstrate their prediction ability.

The remainder of this paper is organized as follows. In Section 2, we describe the link prediction problem in directed networks and the standard metric for performance evaluation. Then, we present how the adapted indices work in asymmetric networks and access their prediction ability in Sections 3 and 4, respectively. Finally, the conclusion is drawn in Section 5.

2. Problem Description and Evaluation Metric

Given a directed network \(G(V, E)\), where \(V\) and \(E\) are sets of nodes and directed links, respectively, multiple links and self-connections are not allowed. The fundamental task of a link prediction algorithm is to give a rank of all nonobserved links in the set \(U \setminus E\), where \(U\) is the universal set containing all \(|V|(|V|−1)\) possible directed links. The mainstream method is to assign each nonobserved link a score and the one with higher score ranks ahead. The top ranked links are regarded as the most likely missing (or future) interactions.

To evaluate the algorithmic performance, the observed links \(E\) is divided into two parts: the training set \(E^T\) is treated as known information, while the probe set \(E^P\) is used for testing and no information therein is allowed to be used for prediction. Clearly, \(E = E^T \cup E^P\) and \(E^T \cap E^P = \emptyset\).

We adopt metric AUC [17] (area under the receiver operating characteristic curve) to quantify the prediction accuracy. It evaluates the predictor performance according to the whole list and can be interpreted as the probability that a randomly chosen missing link (i.e., a link in \(E^P\)) is given a higher score than a randomly chosen nonexistent link (i.e., a link in \(U \setminus E\)). Among \(n\) independent comparisons, if there are \(n'\) times the missing link having a higher score and \(n''\) times they are the same; the AUC value is

\[
\text{AUC} = \frac{n' + 0.5n''}{n}.
\]

Similarly, the link prediction algorithm can also be used to identify spurious interactions. In this case, the task is changed to score all the observed links in \(E\) and rank them in an ascending order; then the top ranked links are suspected as the most likely spurious interactions. To test the algorithmic performance, \(E^T = E \cup E^P\) is treated as the "hypothetical observed" links, where \(E^P\) is made up of some randomly selected nonexistent links from \(U \setminus E\). Now the AUC represents the probability that a randomly chosen spurious link (i.e., a link in \(E^P\)) is given a lower score than a randomly chosen real link (i.e., a link in \(E\)).

In a word, for both predicting missing links and identifying spurious links cases, the higher the AUC value is, the better the prediction algorithm works.

3. Directed Link Predictors

In this section, we first review some classical prediction indices for undirected network and present how to extend them to directed version.

(1) Preferential Attachment (PA). This is a basic link prediction algorithm corresponding to the preferential attachment phenomena in many real-world networks [18, 19]. Preferential attachment means that the more connected a node is, the more likely it is to receive new links. Nodes with higher degree have stronger ability to grab links added to the network. So PA index defines that the probability that a link will connect \(i\) and \(j\) is proportional to the number of their neighbors; that is,

\[
S_{ij}^{PA} = k_i \times k_j = |\Gamma(\dot{i})| \times |\Gamma(\ddot{j})|,
\]

where \(\Gamma(i) = \{j \mid (i, j) \in E \} \cup \{j \mid (j, i) \in E\}\) is the set of node \(i\)'s neighbors and \(k_i = |\Gamma(i)|\) denotes the degree of it.

In directed networks, taking the link direction into consideration, \(\Gamma_{out}(i) = \{j \mid (i, j) \in E\}\) is regarded as the set of outgoing neighbors, while \(\Gamma_{in}(i) = \{j \mid (j, i) \in E\}\) is the group of incoming neighbors. And \(k_{out}^i = |\Gamma_{out}(i)|\) and \(k_{in}^i = |\Gamma_{in}(i)|\) represents the out-degree and in-degree of node \(i\), respectively. Usually, \(k_{out}^i\) and \(k_{in}^i\) are not equal, so we can take advantage of this asymmetry to extend PA index. For an arc \(i \rightarrow j\) in directed networks, the PA score could be defined as

\[
S_{i \rightarrow j}^{PA} = k_{out}^i \times k_{in}^j = |\Gamma_{out}(\dot{i})| \times |\Gamma_{in}(\ddot{j})|.
\]

Actually, the meaning of (3) is quite straightforward: from the perspective of information dissemination, the more outgoing neighbors of \(i\) imply the more channels to get the message out; the more incoming neighbors \(j\) attracts, the
stronger its ability to gather information, so it is more likely to build a connection from $i$ to $j$.

(2) Common Neighbors (CN). The common neighbors refer to the nodes that are connected with both $i$ and $j$. This concept is very important that nearly all the local similarity measures are based on it.

Triadic closure principle in social network analysis states that if two people A and B have a common friend C, they are very likely to be friends too; thus these three nodes form a closed triangle ABC. For undirected networks, it is natural to assume that the more common neighbors two nodes share, the more similar they are and the more likely to build a connection between themselves. So the CN index is defined as

$$S_{ij}^{CN} = |\Gamma(i) \cap \Gamma(j)|. \quad (4)$$

It is not hard to find that $S_{ij}^{CN} = (A^2)_{ij}$, where $A^2 = A \cdot A$. $A$ is the adjacency matrix, the element of which is defined as

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ is connected with } j \\ 0 & \text{otherwise}, \end{cases} \quad (5)$$

and obviously $A_{ji} = A_{ij}$.

However, in directed networks, adjacency matrix $\overline{A}$ becomes asymmetric; that is,

$$\overline{A}_{ij} = \begin{cases} 1 & \text{if there is a directed arc } i \rightarrow j \\ 0 & \text{otherwise}. \end{cases} \quad (6)$$

Now the meaning of $(\overline{A}^2)_{ij}$ changes, which equals the number of different two-step paths from $i$ to $j$; in other words, $(\overline{A}^2)_{ij}$ measures how many “transit nodes” in the network that could forward the message got from $i$ to $j$. The more numbers of such “transit nodes,” the more likely to build a directed arc $i \rightarrow j$. Thus the extension of CN index in directed networks could be naturally defined as

$$S_{i\rightarrow j}^{CN} = (\overline{A}^2)_{ij} = |\Gamma_{out}(i) \cap \Gamma_{in}(j)|. \quad (7)$$

As shown in Figure 1, from the analysis of geometry, CN indices could be interpreted as the number of triadic closure in undirected networks, while in directed networks this index measures the formation of feedforward structure.

In fact, (7) is not the only way to extend CN index. Other alternative indicators include the following.

(i) Co-citation index: $|\Gamma_{out}(i) \cap \Gamma_{out}(j)|$.

(ii) Bibliographic coupling index: $|\Gamma_{in}(i) \cap \Gamma_{in}(j)|$.

(iii) Feedback structure: $|\Gamma_{in}(i) \cap \Gamma_{out}(j)|$.

From Figure 2 we can see that co-citation index and bibliographic coupling index cannot present different judgement for $i \rightarrow j$ and $j \rightarrow i$. The existence possibility for these two arcs will always be the same, so co-citation index and bibliographic coupling index are not suitable for link prediction in directed networks. And we choose feedforward structure instead of feedback structure as the indicator, because the empirical research demonstrates that the former one is more popular in real-world networks. Milo et al. [20] studied the network motifs in networks from biochemistry, neurobiology, ecology, and engineering. They found that feedforward structure appears in all the observed network.

(3) Stochastic Block Model (BM) [14]. Stochastic block model is one of the most general network models, where nodes are partitioned into groups and the probability that two nodes are connected depends only on the groups to which they belong.

In undirected networks, given a block model $M = (P, Q)$, in which $P$ is the partition of nodes into groups and the matrix $Q = (Q_{\alpha\beta})$ describes the probabilities of the linkage between groups, then the likelihood of the observed network is

$$p(A | P, Q) = \prod_{\alpha\beta} Q_{\alpha\beta}^{I_{\alpha\beta}} (1 - Q_{\alpha\beta})^{Y_{\alpha\beta} - I_{\alpha\beta}}, \quad (8)$$

where $A$ is the adjacency matrix, $I_{\alpha\beta}$ is the number of links between nodes in groups $\alpha$ and $\beta$, and $Y_{\alpha\beta}$ is the maximum number of such links. Using Bayes theorem, the reliability of an individual link is

$$S_{ij}^{BM} = p(A_{ij} = 1 | A)$$

$$= \frac{1}{Z} \sum_{P \in \mathcal{P}} \int_{[0,1]^2} p(A_{ij} = 1 | P, Q) p \left( (P, Q) \right) dQ,$$

where $\mathcal{P}$ is the space of all partitions, $G$ is the number of distinct group pairs, and $Z$ is a normalizing constant ($Z = \sum_{P \in \mathcal{P}} \int_{[0,1]^2} p(A | P, Q) p(P, Q) dQ$).

In directed networks, the interaction between two nodes are no longer reciprocal, which implies that the underlying
matrix $Q$ becomes asymmetric; that is, $Q_{\alpha\beta}$ determines the probability of nodes in group $\alpha$ linking to the ones in group $\beta$, while $Q_{\beta\alpha}$ limits the chance of links from group $\beta$ to $\alpha$. $Q_{\alpha\beta}$ and $Q_{\beta\alpha}$ are not necessarily the same. In addition, the links of different direction between two groups should be counted separately. Thus, the likelihood of the directed network structure is

$$p^* (A | P, Q) = \prod_{\alpha, \beta} Q_{\alpha\beta}^\alpha (1 - Q_{\alpha\beta})^{1 - Q_{\alpha\beta}}.$$  \hspace{1cm} (10)

Notice that $I^*_{\alpha\beta}$ now becomes the number of directed arcs from groups $\alpha$ to $\beta$, and $\gamma^*_{\alpha\beta}$ is the corresponding maximum number of such links. Consider

$$\gamma^*_{\alpha\beta} = \begin{cases} |\alpha| - 1 & \alpha = \beta \\ |\alpha| |\beta| & \alpha \neq \beta \end{cases}.$$  \hspace{1cm} (11)

Then by replacing $p(A | P, Q)$ with $p^* (A | P, Q)$ in (9), we can easily get the score for the arc $i \rightarrow j$: $\sum_{p \in \mathcal{P}} \int_{[0,1]^2} p^* (A | P, Q) p(P, Q) dQ$.

$$\delta^B_{i \rightarrow j} = \frac{1}{Z^*} \sum_{p \in \mathcal{P}} \int_{[0,1]^2} p^* (A_{ij} = 1 | P, Q) p^* (A | P, Q) p(P, Q) dQ.$$  \hspace{1cm} (12)

Among all the link prediction measures for undirected networks, the above three indices are chosen to be adapted for the following reasons. Firstly, the design philosophies of them are different from each other, and they all have their own advantages. CN index reflects the mechanism of rich-get-richer and is superior for its least information requirement, while BM method takes into account network community structures, rests on solid mathematical foundations, and returns excellent results. Secondly, those three indices are fairly representative, so their modification methods are enlightening and could be extended to other indices. For example, inspired by the modified CN index, other local similarity measures can also be extended in a similar way. Table 1 illustrates a few common local similarity indices and their corresponding extensions.

To analyze the performance of those modified measures, next we introduce bifan predictor as a comparison index.

![Figure 2: Alternative indicators for CN index.](image)

**Figure 2:** Alternative indicators for CN index.

| Index            | Undirected network | Directed network |
|------------------|--------------------|------------------|
| Jaccard          | $|\Gamma(i) \cap \Gamma(j)|$ | $|\Gamma_{\text{out}}(i) \cap \Gamma_{\text{in}}(j)|$ |
|                  | $|\Gamma(i) \cup \Gamma(j)|$ | $|\Gamma_{\text{out}}(i) \cup \Gamma_{\text{in}}(j)|$ |
| Hub Promoted     | $\min(|k(i), k(j)|)$ | $\min(k_{\text{out}}(i), k_{\text{in}}(j))$ |
| Hub Depressed    | $\max(k(i), k(j))$ | $\max(k_{\text{out}}(i), k_{\text{in}}(j))$ |
| Salton           | $\sqrt{k(i) \times k(j)}$ | $\sqrt{k_{\text{out}}(i) \times k_{\text{in}}(j)}$ |
| Leicht-Holme-Newman | $\frac{|k(i) \times k(j)|}{k_{\text{out}}(i) \times k_{\text{in}}(j)}$ | $\frac{|k_{\text{out}}(i) \times k_{\text{in}}(j)|}{k_{\text{out}}(i) \times k_{\text{in}}(j)}$ |
| Sorensen         | $\frac{2|k(i) \times k(j)|}{k(i) + k(j)}$ | $\frac{2|k_{\text{out}}(i) \times k_{\text{in}}(j)|}{k_{\text{out}}(i) + k_{\text{in}}(j)}$ |
| Adamic-Adar      | $\sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log(k(z))}$ | $\sum_{z \in \Gamma_{\text{out}}(i) \cap \Gamma_{\text{in}}(j)} \frac{1}{\log(k_{\text{out}}(z))}$ |
| Resource Allocation | $\sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k(z)}$ | $\sum_{z \in \Gamma_{\text{out}}(i) \cap \Gamma_{\text{in}}(j)} \frac{1}{k_{\text{out}}(z)}$ |

**Table 1:** Some local similarity measures and their corresponding extensions for directed networks.

Bifan structure is found to be quite widespread in real-world [20] and proven to be a well-performed predictor for directed networks.

(4) **Bifan Predictor** [16]. The potential theory is proposed as a microscopic organizing principle for directed networks, which assumes that each directed link corresponds to a decrease of a unit potential and subgraphs with definable potential values for all nodes are preferred. Combining the clustering and homophily mechanisms with potential theory, it is deduced that bifan subgraph is the most favored local structure in directed networks (see Figure 3).

This special structure could be directly used as a link predictor, with the assumption that a link that can generate more bifan subgraphs is more significant and is thus of a higher probability to appear. So the bifan index is determined by how many bifan subgraphs that link $i \rightarrow j$ could generate, which can be calculated by

$$\delta^\text{Bifan}_{i \rightarrow j} = r_i \left( \sum_{k} r_{k}^{T} \right)$$  \hspace{1cm} (1)
Comparisons of algorithms’ accuracy are displayed in Figures 4 and 5. The x-axis denotes fraction $f = |E^f|/|E|$ for missing interactions ($f = |E^p|/|E|$ for spurious interactions), ranging from 0.05 to 0.95, and the interval is set to be 0.05. The dashed line indicates the baseline accuracy when the score of each link is got by pure chance. Each AUC value is obtained by averaging 100 independent realizations, except for the PB, of which the implementation times are 50.

4.2.1. Comparison of Local Similarity Measures. Figure 4 illustrates the experimental results for 10 extended local similarity indices. It is worth noting that PA index, which is regarded as the worst predictor in undirected networks, surprisingly performs quite well in the experiments. It outperforms all the other local similarity measures in identifying both missing and spurious links. Hence the relatively good prediction ability together with the least information requirement makes PA index more competitive in directed networks.

Note that when applied in finding missing links, the performance of other nine local similarity measures based on common neighbors is nearly the same, especially in CE and PB networks. It is because all these extended indices are on the basis of feedforward structure, if and only if there are some “transit nodes” between i and j; the score of arc $i \rightarrow j$ is larger than 0. So when the average shortest distance $\langle d \rangle$ is relatively large, the score of lots of arcs will be zero for all these measures, which leads to the little difference in their prediction ability.

4.2.2. Comparison of Different Kinds of Measures. In Figure 5, we choose PA and CN as representatives of local similarity measures and compare their performance with BM and bifan. It is obvious that when applied in finding missing links, BM outperforms CN and PA and works even better than bifan predictor in most networks. The advantage of BM to others is usually remarkable except for network PB, in which the performances of BM, bifan, and PA nearly keep pace with each other. This outcome demonstrates that community structure also plays an important role in directed networks. Compared with other three measures utilizing only the local structure information, BM index which takes advantage of the organizing principles of the whole network could get more accurate prediction results.

Figure 5 also shows that the prediction accuracy of all four indices decline with the increases of $f$, but the rate of decrease is varying. Compared with the linear decrease of CN index, the trend of AUC for others is slowly downward until it reaches a “turning point,” after which the remaining links in $E^f$ are not enough to infer the underlying structure of network, so the prediction accuracy drops sharply. For example, in all FoodWebs, the prediction accuracy falls dramatically after $f$ reaches 0.75. In network PB, 0.65 could be regarded as the “turning point,” before which the AUC values for PA, BM, and bifan nearly remain unchanged, but after that the prediction accuracy drops quickly. Note that in all five networks the AUC of bifan is even worse than PA when the value of $f$ exceeds some value. Bifan index is defined as the number of bifan subgraphs that link $i \rightarrow j$ could.

4.1. Data Sets. Our experiments are carried out on five real-world directed networks drawn from biological and technological fields. Those networks are pretty representative and are often employed in simulation experiments to validate the effectiveness of the proposed scheme. (i) FoodWeb1 (FW1) [21]: a network of food web represents the predator-prey relations between 69 species living in Everglades Graminoids during wet season. (ii) FoodWeb2 (FW2) [22]: a network of food web consists of 97 species living in Florida Bay during wet season. (iii) FoodWeb3 (FW3) [23]: a network of food web consists of 128 species living in Florida Bay during wet season. (iv) C. elegans (CE) [24]: it is the neural network of the nematode worm C. elegans, in which an edge joins two neurons if they are connected by either a synapse or a gap junction. (vi) Political blogs (PB) [25]: it is a directed networks of hyperlinks between weblogs on US political blogs.

For convenience, we eliminate all the loops and multi-edges of the above networks. The basic topological features of these five networks are summarized in Table 2.

4.2. Experimental Results. In this section, we investigate the prediction ability of the indices presented in Section 3.
Figure 4: Continued.
generate. With the increase of $f$, more and more links are removed from $E'$, and it is more and more difficult to build bifan structure, which leads to the decline of discrimination for $S_{i \rightarrow j}^{Bifan}$. While the increase of $f$ does not affect the PA index that much, so the prediction accuracy of bifan is lower than PA when lots of the links are removed.

As is illustrated in Figure 5, in identifying spurious interactions, BM and bifan are the two best predictors which are followed by PA, while CN index still has the worst overall performance as before. Note that, for each index, different from the results in finding missing links, AUC values are relatively high and are basically not changed with the increases of $f$. However, high accuracy is not sufficient for spurious link detection algorithms. If just a few unexpected important links are incorrectly removed, the structural and dynamical properties of the network may change dramatically [26]. So when applied in identifying spurious links, the prediction indices should be evaluated meticulously; however, this problem is beyond the scope of this paper.

5. Conclusion

This paper studied how to identify both missing and spurious interactions in directed networks. We have showed how to extend classical link prediction indices to directed version, making them able to predict both the existence and direction of an arc between two nodes. Simulation experiments on five real-world directed networks have vividly demonstrated the prediction ability of those modified measures.
Figure 5: Continued.
The purpose of this paper is not to present a better directed link predictor, but to give an example of how to use the valuable knowledge of undirected networks to solve the problems in directed networks. For directed networks, the direction of links is a double-edged sword. It provides additional valuable information for link prediction but also leads to greater difficulty—how to determine the direction of the unknown links is a tough problem. The extended CN index made use of the asymmetric similarity to determine the link direction, while BM benefited from asymmetric likelihood of the linkage between groups. In brief, asymmetry is the key. However, prediction methods based solely on asymmetry are incomplete, and other natures of the directed networks should be taken into consideration. For example, in neural networks and some technological networks, information or resources are often collected by basal nodes, transmitted through directed links, and delivered to top nodes eventually. Such kind of macroscopic flow direction may help to determine the unknown link direction.

On the other hand, each prediction index has its own advantages and is suitable for different kind of networks. How to choose the most appropriate algorithm according to the features of the networks is also a problem worth study.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The work is supported by Basic Research Project of National University of Defense Technology and the Open Fund from HPCL no. 201403-01.

References

[1] Y. Lin, X. Duan, C. Zhao, and L. Xu, System Science: Methodological Approaches, CRC Press, Boca Raton, Fla, USA, 2013.
[2] W. Cheng, X. Cheng, T. Znati, X. Lu, and Z. Lu, “The complexity of channel scheduling in multi-radio multi-channel wireless networks,” in Proceedings of the IEEE INFOCOM, pp. 1512–1520, Rio de Janeiro, Brazil, April 2009.
[3] J. L. Schafer and J. W. Graham, “Missing data: our view of the state of the art,” Psychological Methods, vol. 7, no. 2, pp. 147–177, 2002.
[4] G. Kossinets, “Effects of missing data in social networks,” Social Networks, vol. 28, no. 3, pp. 247–268, 2006.
[5] 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.
[6] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, “Recommender systems,” Physics Reports, vol. 519, no. 1, pp. 1–49, 2012.
[7] A. Clauset, C. Moore, and M. E. J. Newman, “Hierarchical structure and the prediction of missing links in networks,” Nature, vol. 453, no. 7191, pp. 98–101, 2008.
[8] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” Journal of the American Society for Information Science and Technology, vol. 58, no. 7, pp. 1019–1031, 2007.
[9] M. E. J. Newman, “Clustering and preferential attachment in growing networks,” Physical Review E, vol. 64, no. 2, Article ID 025102, 2001.
[10] P. Jaccard, “Etude comparative de la distribution orale dans une portion des Alpeset des Jura,” Bulletin de la Société Vaudoise des Sciences Naturelles, vol. 37, pp. 547–579, 1901.
[11] L. A. Adamic and E. Adar, “Friends and neighbors on the Web,” Social Networks, vol. 25, no. 3, pp. 211–230, 2003.
[12] T. Zhou, L. Lü, and Y.-C. Zhang, “Predicting missing links via local information,” European Physical Journal B, vol. 71, no. 4, pp. 623–630, 2009.
[13] X. Feng, J. C. Zhao, and K. Xu, “Link prediction in complex networks: a clustering perspective,” *European Physical Journal B*, vol. 85, no. 1, article 3, 2012.

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

[15] X. Zhang, X. Wang, C. Zhao, D. Yi, and Z. Xie, “Degree-corrected stochastic block models and reliability in networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 393, pp. 553–559, 2014.

[16] 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.

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

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

[19] Y.-B. Xie, T. Zhou, and B.-H. Wang, “Scale-free networks without growth,” *Physica A*, vol. 387, no. 7, pp. 1683–1688, 2008.

[20] R. Milo, S. Shen-Orr, N. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks,” *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[21] R. E. Ulanowicz, J. J. Heymans, and M. S. Egnotovich, “Network analysis of trophic dynamics in South Florida ecosystems, FY 99: the graminoid ecosystem,” Tech. Rep. TS-191-99, 2000.

[22] D. Baird, J. Luczkovich, and R. R. Christian, “Assessment of spatial and temporal variability in ecosystem attributes of the St Marks national wildlife refuge, Apalachee Bay, Florida,” *Estuarine, Coastal and Shelf Science*, vol. 47, no. 3, pp. 329–349, 1998.

[23] R. E. Ulanowicz, C. Bondavalli, and M. S. Egnotovich, “Network analysis of trophic dynamics in South Florida ecosystem, FY 97,” Tech. Rep. CBL 98-123, The Florida Bay Ecosystem, 1998.

[24] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, “The structure of the nervous system of the nematode Caenorhabditis elegans,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 314, no. 1165, pp. 1–340, 1986.

[25] A. Adamic and N. Glance, “The political blogosphere and the 2004 US election,” in *Proceedings of the Workshop on the Weblogging Ecosystem (WWW ’05)*, 2005.

[26] A. Zeng and G. Cimini, “Removing spurious interactions in complex networks,” *Physical Review E*, vol. 85, no. 3, Article ID 036101, 2012.