IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 24, 2019, accepted December 17, 2019, date of publication December 23, 2019,
date of current version January 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961399

Link Prediction by Multiple Motifs in Directed Networks

YAFANG LIU, TING LI, AND XIAOKE XU, (Member, IEEE)
College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China

Corresponding author: Xiaoke Xu (xuxiaoke@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61603073, Grant 61601081, and Grant 61773091, in part by the Key Research and Development Plan of Liaoning Province under Grant 2018104016, in part by the Liaoning Revitalization Talents Program under Grant XLYC1807106, and in part by the Program for the Outstanding Innovative Talents of Higher Learning Institutions of Liaoning under Grant LR2016070.

ABSTRACT Link prediction which can restore and predict missing links has wide applications in complex networks. In existing researches on link prediction of directed networks, most methods only consider the information of a single motif, in which the effects of multiple motifs are not included, especially the role of each node forming different motif structures. In order to solve the above problems, firstly we propose a single motif naive Bayes model beyond calculating the number of edge-dependent motifs. We also investigate a two-motif naive Bayes model and a machine learning framework based on multi-motif features to further improve the performance of link prediction. The new framework of link prediction by multiple motifs is superior to the state-of-the-art methods such as potential theory, local path and superposed random walk. Experimental results on real-life networks show the highest performance improvement is 64.3%. Finally, we use maximal information coefficients to reveal the topology correlation between different motifs, which is helpful to understand the evolutionary mechanism of directed networks.

INDEX TERMS Link prediction, motif, naive Bayes, directed networks.

I. INTRODUCTION

Link prediction is a hot topic in the field of complex networks [1], [2], and its basic idea is to predict the possibility of a link existing between two nodes using partial information of network structures [3], [4]. Link prediction can uncover the probability of a link appearing in future in a dynamic network [5], [6], and it can find out whether an observed link is false or not in a static network [7], [8].

In the methods of link prediction based on local topology structures, the indicator of common neighbor similarity is the most commonly used [9], [10]. However, this kind of methods does not consider the direction of links, so they cannot be applied directly to directed networks. In a real-life network, the relationship between a pair of individuals is often asymmetric. A network that uses direction information to represent the asymmetry relationship is called a directed network. Currently, many real-world networks can be described as directed networks, such as online social services [11], food chains [12], and power delivery systems [13].

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

By adding directions to the connected subgraphs composed of several nodes, the local structures of a directed network can be constructed. Although there are many indicators to describe the structural characteristics of a directed network, one of the most significant methods is motif analysis [14], [15]. Milo [16] refers to a connected subgraph that the number of it appearing in a real-world network is significantly higher than that in the corresponding randomized networks. Currently, only a few studies have used motif theory for link prediction in directed networks. Zhang *et al.* proposed potential theory of directed networks, and found the motifs that satisfied potential theory had a higher accuracy for link prediction, especially the motif of Bi-fan [17]. However, only the prediction based on definable potential motifs were considered, and the other types of motifs have not been considered yet. Furthermore, in each algorithm only one single motif is considered for link prediction, without taking the joint effects of multiple motifs into account.

In the method of potential theory and the indices of local structure similarity, the researchers assumed that the nodes in the neighborhood of the predicted edge contributed the same to the composition of motifs. However, in a lot of real-life network systems, this idea might not be correct.

Actually, if we distinguish the different contribution of each node using a role function induced by a local naive Bayes model, better prediction results can be obtained. For instance, Liu *et al*. used the naive Bayes model and the indices of common neighbors to predict the unobserved links in undirected networks, and the experimental results showed that the prediction performance of most real-life networks can be improved [18]. Moreover, in the study of link prediction for weighted networks, Wu et al. proposed a probabilistic prediction framework based on weighted local naive Bayes model, and the experimental results suggested that the accuracy of link prediction can be improved compared with traditional algorithms [19]. However, in existing researches on link prediction of directed networks, the different roles of each node have not been considered yet.

To solve the above problems, in this study we propose a link prediction method for directed networks based on multi-motif information. Firstly we propose a single motif naive Bayes model instead of only calculating the number of edge-dependent motifs. Next, we construct a two-motif naive Bayes model and a multi-motif based link prediction method using a machine learning framework. The experimental results suggest that the proposed method is superior to the state-of-the-art methods, such as potential theory, resource allocation, local path, and superposed random walk. At last, the correlation between the motif predictors can verify the rationality of calculating the role function of 4-node motifs by analyzing the predictor structure of the same type.

The information in each section of this study is as follows. In Section II, we introduce the experimental directed networks and evaluation indicators. In Section III, a single-motif based naive Bayes model is proposed to perform the task of link prediction. In Section IV, the two-motif naive Bayes model is proposed. In Section V, a multi-motif based link prediction method with a framework of machine learning is proposed. In the end, we conclude this study.

II. EXPERIMENTAL DATA AND EVALUATION INDICATOR

A. DATA DESCRIPTION

In this study, we use seven real-life networks to verify the performance of different methods for link prediction. Several statistics of the experimental networks are shown in Table 1.

Bison [20]–[22] - This directed network contains dominance between American bisons in 1972 on the National Bison Range in Moiese. A node represents a bison and an edge represents dominance of the left bison over the right bison. This network has 26 nodes and 314 links.

FWEG [23] - A Food Web living in Everglades Graminoids during wet season (FWEG), contains 69 species of creatures and 916 predator relationships.

Macaques [22], [24], [25] - This directed network contains 1187 dominance behaviors in a colony of 62 adult female Japanese macaques.

TABLE 1. Description of the experimental network data. m and n are the number of nodes and links, k_{max} and $<k>$ are the maximum degree and the average degree of all nodes, r represents the assortativity coefficient, c is the average clustering coefficient, and l is the average shortest path length.

Network	m	n	$<k>$	k_{max}	r	c	l
Bison	26	314	17.07	24	-0.10	0.81	1.32
FWEG	69	916	25.65	64	-0.29	0.55	1.64
Macaques	62	1187	37.65	55	-0.07	0.67	1.38
FWME	97	1492	29.81	90	-0.15	0.47	1.69
CESF	128	2137	32.91	110	-0.10	0.34	1.77
C. elegans	297	2345	14.47	134	-0.16	0.29	2.46
SmaGri	1024	4919	9.60	232	-0.19	0.31	2.98

FWME [26] - A Food Web living in Mangrove Estuary during wet season (FWME), contains 97 species of creatures and 1492 predator relationships.

CESF [22], [23], [27] - This network contains 2137 Carbon Exchanges in the cypress wetlands of South Florida during the dry season (CESF). Nodes represent taxa and an edge denotes that a taxon uses another taxon as food with a given trophic factor.

C. elegans [28] - A 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. This network has 297 nodes and 2345 links.

SmaGri [29] - The citations of Small & Griffith and descendants (SmaGri). This network contains 1024 nodes and 4919 links.

B. EVALUATION INDICATOR

The AUC value is defined as the area under the receiver operating characteristic curve [3], which is used to quantify the performance of link prediction here. Given a set of missing links and a set of non-existing edges with the same length, AUC can be referring as the probability that the scores of missing edges are higher than those of non-existing edges. A link is selected from the missing set and the nonexistent set respectively. If the link from the missing set has a higher score, the value of AUC adds 1 point; if the links from the missing set and non-existing set have the same score, the value of AUC adds 0.5 points; otherwise no extra points will be added.

The comparisons are performed n times independently. There are X comparisons where the score of the missing link is higher than that of the non-existing link, and there are Y comparisons where the score of the missing link equals to that of the non-existing link, then AUC can be defined as:

$$AUC = \frac{X + 0.5Y}{n}. \tag{1}$$

The results of AUC are generally between 0.5 and 1. The larger the value of AUC is, the better the performance of the predicted algorithm is.

III. LINK PREDICTION BY SINGLE MOTIF NAIVE BAYES MODEL

A. LINK PREDICTION BY EDGE-DEPENDENT MOTIF

A directed network refers to a network that has an unequal relationship between each pair of nodes in a real complex system. For directed networks, Zhang *et al*. proposed a motif classification method called potential theory [17]. Given a subgraph, only if each node in the subgraph can be assigned a potential, it is definable. The condition that nodes can be assigned potential energy is: for each pair of nodes v_i and v_j, if $v_i \rightarrow v_j$, the potential energy of v_i is 1 unit higher than v_j. If a subgraph contains a reciprocal link, it is not potential-definable. The authors found the definable motifs had higher performance for link prediction than undefinable motifs, especially the Bi-fan motif. However, in their study only one single motif is considered for each method of link prediction, without fusing the joint effects of multiple motifs.

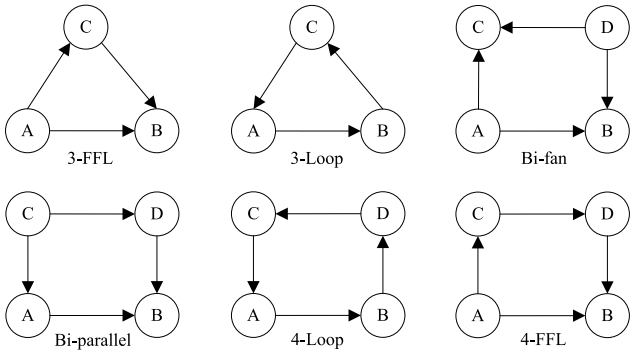

FIGURE 1. 3-node and 4-node connected subgraphs without reciprocal links in directed networks.

To make comprehensive use of all types of motif information, we list all the 3-node and 4-node subgraphs with loops, shown in Fig. 1. For the six kinds of subgraphs, only Bi-fan and Bi-parallel motifs are definable potentials, and the others are newly constructed motifs that can be utilized for link prediction. We only choose the motifs with three and four nodes instead of higher-order motifs for two reasons. First, the fewer the number of nodes in a motif is, the easier it is to calculate. Second, the number of the higher-order motifs depends on the number of lower-order motifs [30], so the motifs with five nodes (5th-order) and above do not play a leading role in link prediction [31]. In order to compare the proposed method with the method of potential theory, we only chose the subgraphs without reciprocal links.

By selecting one link from each motif as the predicted link, a motif predictor for link prediction can be constructed. Based on the six types of motifs in Fig. 1, 12 predictors can be obtained. As shown in Fig. 2, the red dotted line in each predictor represents the predicted link. The method of potential theory respects that if the existence of the link to be predicted can generate more definable potential motifs, the possibility of this link appearing is greater. In this theory, only the prediction based on definable potential motifs is considered. In the motif predictors shown in Fig. 2, the number

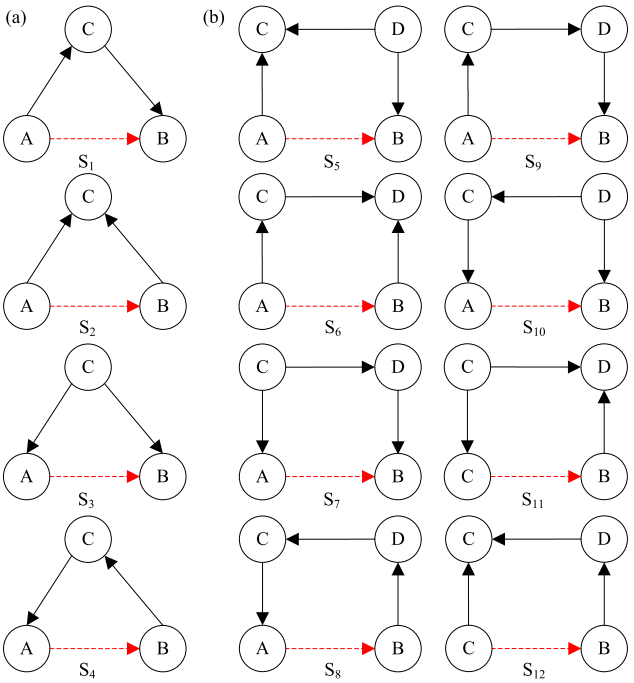

FIGURE 2. 12 predictors based on the 3-node and 4-node motifs. (a) The predictors of 3-node motifs, and (b) the predictors of 4-node motifs.

of motifs containing the predicted link can be calculated, which is called as the predictor based on the number of edge-dependent motifs. From this point of view, potential theory can be regarded as a special case of link prediction by calculating the number of edge-dependent motifs, because it only considers the motifs that satisfy definable potentials. Therefore, the method based on the number of edge-dependent motifs can use the structure information of all types of motifs more comprehensively.

B. THE ROLE FUNCTION OF DIFFERENT NODES

The algorithm based on edge-dependent motifs considers the number of motifs on the predicted edge as a key indicator for link prediction. It is believed that the more the number of specified motifs that are composed of a pair of nodes with other nodes (3-node motifs) or edges (4-node motifs) is, the more likely this node pair is to be connected. In this method, it is assumed that the nodes in the neighborhood of the predicted edge contribute the same to the composition of motifs, though this idea might not be always correct in real-life networks. Here, we argue that the nodes and edges in the neighborhood of the predicted edge play different roles and thus lead to distinct contributions, and propose a motif based naive Bayes model. The model can effectively distinguish the contributions of different nodes and edges that constitute the corresponding motif, and achieve better performance for link prediction.

Here, we use a small toy network as shown in Fig. 3(a) to illustrate the different contributions of nodes in the neighborhood of the predicted edge. Taking predictor S_1 as an example, two nodes (C_1 and C_2) can form predictor S_1 with the predicted edge (A, B). If we only consider the number

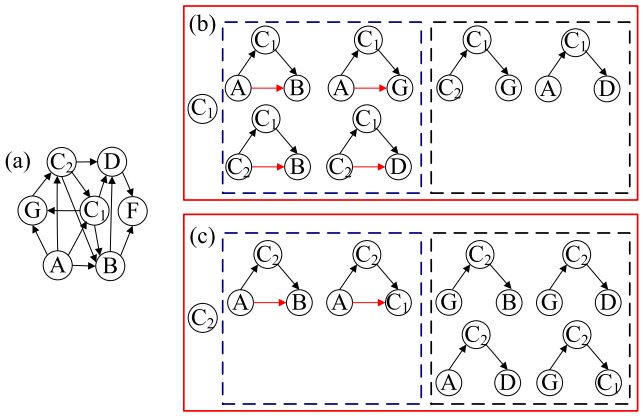

FIGURE 3. Analysis of the role functions of nodes C_1 and C_2. (a) A small toy network, (b) the role function of node C_1, and (c) the role function of node C_2.

of edge-dependent motifs, the contributions of C_1 and C_2 are the same. However, as shown in Fig. 3(b), node C_1 can form four predictor S_1 with other node pairs (as shown in the blue box) and two pairs of nodes that are not connected (as shown in the black box). Therefore, the probability of the connection between A and B that combines S_1 with C_1 is $4/(4+2)$=66.7%. As shown in Fig. 3(c), node C_2 can form predictor S_1 with two connected node pairs (as shown in the blue box) and four pairs of nodes that are not connected (as shown in the black box), so the probability of connection between A and B that combines S_1 with C_2 is $2/(2+4)$=33.3%. In this case, considering the number of edge-dependent motifs only and ignoring the distinguishing roles of C_1 and C_2 lead to the performance of link prediction inaccurate. Therefore, in the following we will propose a Single Motif Naive Bayes (SMNB) model to consider not only the number of edge-dependent motifs but also the contributions of different nodes.

C. SINGLE MOTIF NAIVE BAYES MODEL

First, we calculate the priori probability whether there is a link between a pair of nodes x and y. Given a network $G(V, E)$ and a training set E^T, the class variable e_{xy} denotes a pair of nodes which are connected, and the class variable $\overline{e_{xy}}$ denotes disconnection. Then the priori probability $P(e_{xy})$ and $P(\overline{e_{xy}})$ can be calculated as

$$P(e_{xy}) = \frac{M}{M^F}, \tag{2}$$

$$P(\overline{e_{xy}}) = \frac{M^F - M}{M^F}, \tag{3}$$

where M^F=$|V|(|V|-1)$, which indicates the number of all the possible links in a network; and M=$|E^T|$, which indicates the number of missing links. Node w owns two conditional probabilities $P(w|e_{xy})$ and $P(w|\overline{e_{xy}})$. $P(w|e_{xy})$ indicates the probability that node w participates in combining the certain predictor when nodes x and y are connected. $P(w|\overline{e_{xy}})$ indicates the probability that node w participates in combining the certain predictor when nodes x and y are disconnected. According to Bayes' theorem, the two conditional probabilities can be calculated as

$$P(w|e_{xy}) = \frac{P(w) \cdot P(e_{xy}|w)}{P(e_{xy})}, \tag{4}$$

$$P(w|\overline{e_{xy}}) = \frac{P(w) \cdot P(\overline{e_{xy}}|w)}{P(\overline{e_{xy}})}, \tag{5}$$

where $P(w)$ indicates the probability that node w and a pair of nodes can constitute a certain predictor. For a pair of nodes (x, y), $O(x, y)$ is the set of nodes that can form a certain predictor with them. Treat the nodes in set $O(x, y)$ as feature variables and assume that these feature variables are independent of each other. According to naive Bayes theory, the posterior probability that nodes x and y are connected or disconnected to each other can be calculated as

$$\begin{aligned} P(e_{xy}|O(x,y)) &= \frac{P(e_{xy})}{P(O(x,y))} \cdot P(O(x,y)|e_{xy}) \\ &= \frac{P(e_{xy})}{P(O(x,y))} \cdot \prod_{w \in O(x,y)} P(w|e_{xy}), \end{aligned} \tag{6}$$

$$\begin{aligned} P(\overline{e_{xy}}|O(x,y)) &= \frac{P(\overline{e_{xy}})}{P(O(x,y))} \cdot P(O(x,y)|\overline{e_{xy}}) \\ &= \frac{P(\overline{e_{xy}})}{P(O(x,y))} \cdot \prod_{w \in O(x,y)} P(w|\overline{e_{xy}}). \end{aligned} \tag{7}$$

For a given pair of nodes, we can judge whether they are more inclined to link by comparing the probability $P(e_{xy}|O(x, y))$ that they tend to link and the probability $P(\overline{e_{xy}}|O(x, y))$ that they do not tend to link. To determine which condition is more likely to occur, we calculate the ratio of the two probabilities as

$$r_{xy} = \frac{P(e_{xy})}{P(\overline{e_{xy}})} \cdot \prod_{w \in O(x,y)} \frac{P(\overline{e_{xy}}) \cdot P(e_{xy}|w)}{P(e_{xy}) \cdot P(\overline{e_{xy}}|w)}, \tag{8}$$

where $P(e_{xy}|w)$ indicates the probability of connection between a pair of nodes x and y that can combine a predictor with node w, and $P(\overline{e_{xy}}|w)$ represents the probability of disconnection. We find out the node pair of x and y that form a certain predictor with node w, and then judge whether the link between x and y is connected or not. Therefore, there is

$$P(e_{xy}|w) = \frac{N_{\triangle w}}{N_{\triangle w} + N_{\wedge w}}, \tag{9}$$

where $N_{\triangle w}$ and $N_{\wedge w}$ are the number of connected and disconnected pairs of nodes in a certain predictor with node w respectively. Due to $P(e_{xy}|w)$+$P(\overline{e_{xy}}|w)$= 1, there is

$$\begin{aligned} P(\overline{e_{xy}}|w) &= 1 - P(e_{xy}|w) \\ &= \frac{N_{\wedge w}}{N_{\triangle w} + N_{\wedge w}}. \end{aligned} \tag{10}$$

Let s=$\frac{P(\overline{e_{xy}})}{P(e_{xy})}$, and substitute Eqs. 2 and 3, $\frac{P(\overline{e_{xy}})}{P(e_{xy})}$ can be simplified as

$$\frac{P(\overline{e_{xy}})}{P(e_{xy})} = \frac{M^F - M}{M^F} \cdot \frac{M^F}{M} = \frac{M^F - M}{M}. \tag{11}$$

From Eqs. 8-11, r_{xy} can be simplified as

$$r_{xy} = s^{-1} \prod_{w \in O_{xy}} s \frac{N_{\triangle w}}{N_{\triangle w} + N_{\wedge w}} \cdot \frac{N_{\triangle w} + N_{\wedge w}}{N_{\wedge w}} = s^{-1} \prod_{w \in O_{xy}} s \frac{N_{\triangle w} + 1}{N_{\wedge w} + 1}, \quad (12)$$

where $s = \frac{P(\overline{e_{xy}})}{P(e_{xy})} = \frac{M^F - M}{M}$. In order to prevent the denominator from being 0, the numerator and denominator in the equation should be added by 1. At this point, given a node w, its role function can be defined as

$$R_w = \frac{N_{\triangle w} + 1}{N_{\wedge w} + 1}. \quad (13)$$

Therefore, Eq. 12 can be simplified as

$$r_{xy} = s^{-1} \prod_{w \in O_{xy}} s R_w. \quad (14)$$

The model described above is called Single Motif Naive Bayes (SMNB) model. Obviously, given a network, the value of s can be determined. Take the logarithm of Eq. 14,

$$r'_{xy} = \underbrace{|O_{xy}| \, logs}_{\text{the number of predictor } S_i} + \underbrace{\sum_{w \in O_{xy}} logR_w}_{\text{the role function of predictor } S_i}. \quad (15)$$

r'_{xy} indicates the feature score between nodes x and y, and we can judge the connection possibility based on Eq. 15. It consists of two parts, which form the naive Bayes model of a single motif together. The first part refers to the number of edge-dependent motifs that contain the predicted link, and the second part represents the total contribution of the role function of all the nodes (or edges) that can form the corresponding motif with the node pair of $\{x, y\}$.

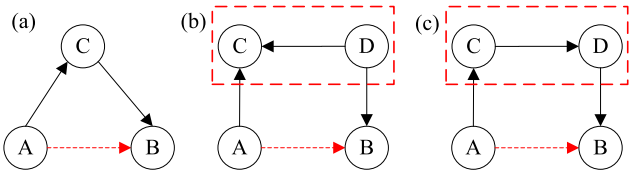

FIGURE 4. The role function calculation of 3-node motif and 4-node motifs.

The role function is based on 3-node motifs, so it cannot be applied to the cases of 4-node motifs directly. As shown in Fig. 4(a), the role function of a 3-node motif considers the influence of node C only. For a 4-node motif, in addition to the predicted link between nodes A and B, there are two additional nodes C and D as shown in Fig. 4(b). There are three possible ways of calculating the role function: (i) only consider the impact of node C; (ii) only consider the impact of node D; (iii) take nodes C and D and the connection between them as a whole, and consider the impact of the whole structure. The first two methods only consider the partial structures of the motif besides the predicted link, so the effect of 4-node motif on link prediction is not revealed. In this study, we use the third way to calculate the role function of 4-node motif. As shown in Figs. 4(b) and (c), we regard the edge structure in the red box as a whole, which is used to calculate the role function for each 4-node motif.

D. PERFORMANCE EVALUATION

In order to verify whether the role function is valid, we compare the performance of SMNB method and the number of edge-dependent motifs based method on multiple real-life networks. Among them, the experimental result of the C. elegans network is shown in Fig. 5. The red scatter indicates the performance of SMNB, and the green scatter indicates the results obtained by the number of edge-dependent motifs based method. It can be seen that adding role functions can improve the performance of link prediction. Similar experimental results can be obtained in several other networks. Therefore, it is necessary to add the role function for link prediction.

IV. LINK PREDICTION BY TWO-MOTIF NAIVE BAYES MODEL

A. TWO-MOTIF NAIVE BAYES MODEL

The experiments have verified that adding the role function can improve the prediction performance for single motif predictors. However, this method of SMNB does not consider the combined effects of multiple motifs. Here, we propose a naive Bayes model integrating two kinds of motifs. The experimental results suggest that the prediction performance based on two-motif can be improved in most cases compared to that of a single motif.

In the structures of two motifs, for a pair of nodes x and y, $O_1(x, y)$ represents the set of nodes that is combined to the first type of predictor with the link, and $O_2(x, y)$ represents the set of nodes that is combined to the second type of predictor with the link. The posterior probabilities that nodes x and y are connected or disconnected can be calculated as

$$\begin{aligned} &P(e_{xy}|(O_1(x,y), O_2(x,y))) \\ &= \frac{P(e_{xy}) \cdot P((O_1(x,y), O_2(x,y))|e_{xy})}{P(O_1(x,y), O_2(x,y))} \\ &= \frac{P(e_{xy}) \cdot P((O_1(x,y))|e_{xy}) P((O_2(x,y))|e_{xy})}{P((O_1(x,y), O_2(x,y))}, \end{aligned} \quad (16)$$

$$\begin{aligned} &P(\overline{e_{xy}}|(O_1(x,y), O_2(x,y))) \\ &= \frac{P(\overline{e_{xy}}) \cdot P((O_1(x,y), O_2(x,y))|\overline{e_{xy}})}{P(O_1(x,y), O_2(x,y))} \\ &= \frac{P(\overline{e_{xy}}) \cdot P((O_1(x,y))|\overline{e_{xy}}) \cdot P((O_2(x,y))|\overline{e_{xy}})}{P(O_1(x,y), O_2(x,y))}. \end{aligned} \quad (17)$$

Given a pair of nodes x and y, comparing the probability of connection $P(e_{xy}|(O(x,y)))$ with the probability of disconnection $P(\overline{e_{xy}}|(O(x,y)))$, we can determine whether they are more inclined to be connected to each other. The ratio of these two

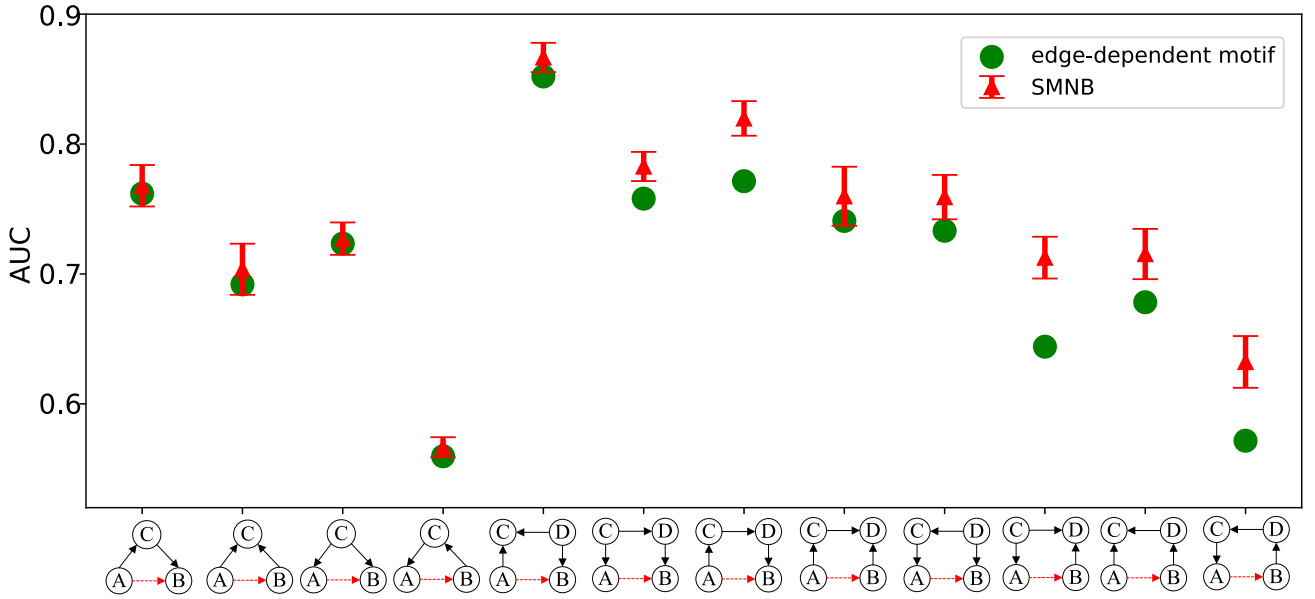

FIGURE 5. Comparison of the prediction performance between the number of edge-dependent motifs and the Single Motif Naive Bayes (SMNB) model.

probabilities can be used to calculate a score for each pair of nodes, which can be calculated as

$$r_{xy} = \frac{P(e_{xy}|(O_1(x,y),O_2(x,y)))}{P(\overline{e_{xy}}|(O_1(x,y),O_2(x,y)))} = \frac{P(e_{xy})}{P(\overline{e_{xy}})} \cdot \frac{P(O_1(x,y)|e_{xy})}{P(O_1(x,y)|\overline{e_{xy}})} \cdot \frac{P(O_2(x,y)|e_{xy})}{P((O_2(x,y))|\overline{e_{xy}})}, \quad (18)$$

where $\frac{P(O_1(x,y)|e_{xy})}{P(O_1(x,y)|\overline{e_{xy}})}$ can be simplified as

$$\frac{P(O_1(x,y)|e_{xy})}{P(O_1(x,y)|\overline{e_{xy}})} = \prod_{w\in O_1(x,y)} \frac{P(w|e_{xy})}{P(w|\overline{e_{xy}})} = \prod_{w\in O_1(x,y)} \frac{P(\overline{e_{xy}})}{P(e_{xy})} \cdot \frac{P(e_{xy}|w)}{P(\overline{e_{xy}}|w)}, \quad (19)$$

where $\frac{P(O_2(x,y)|e_{xy})}{P(O_2(x,y)|\overline{e_{xy}})}$ can be simplified as

$$\frac{P(O_2(x,y)|e_{xy})}{P(O_2(x,y)|\overline{e_{xy}})} = \prod_{v\in O_2(x,y)} \frac{P(v|e_{xy})}{P(v|\overline{e_{xy}})} = \prod_{v\in O_2(x,y)} \frac{P(\overline{e_{xy}})}{P(e_{xy})} \cdot \frac{P(e_{xy}|v)}{P(\overline{e_{xy}}|v)}. \quad (20)$$

Substituting Eqs. 19 and 20, r_{xy} can be rewritten as

$$r_{xy} = \frac{P(e_{xy})}{P(v|\overline{e_{xy}})} \prod_{w\in O_1(x,y)} \frac{P(\overline{e_{xy}})}{P(e_{xy})} \cdot \frac{P(e_{xy}|w)}{P(\overline{e_{xy}}|w)} \prod_{v\in O_2(x,y)} \frac{P(\overline{e_{xy}})}{P(e_{xy})} \cdot \frac{P(e_{xy}|v)}{P(\overline{e_{xy}}|v)}, \quad (21)$$

where $P(e_{xy}|w)$ indicates the probability that node w constitutes closed predictor of the first motif, which can be calculated as

$$P(e_{xy}|w) = R_w = \frac{N_{\triangle w}}{N_{\triangle w} + N_{\wedge w}}. \quad (22)$$

Meanwhile, $P(\overline{e_{xy}}|w)$ indicates the probability that node w constitutes unclosed predictor of the first motif, which can be calculated as

$$P(\overline{e_{xy}}|w) = R_w = \frac{N_{\wedge w}}{N_{\triangle w} + N_{\wedge w}}. \quad (23)$$

In order to prevent the denominator from being 0, the numerator and denominator in the equation are added by 1. Using Eqs. 22 and 23, the role function R_w of node w obtained from the first predictor can be computed as

$$R_w = \frac{P(e_{xy}|w)}{P(\overline{e_{xy}}|w)} = \frac{N_{\triangle w}+1}{N_{\wedge w}+1}. \quad (24)$$

Similarly, the role function R_v of node v for the second predictor can be calculated as

$$R_v = \frac{P(e_{xy}|v)}{P(\overline{e_{xy}}|v)} = \frac{N_{\triangle v}+1}{N_{\wedge v}+1}. \quad (25)$$

Substituting Eqs. 21, 24 and 25, r_{xy} can be simplified as

$$r_{xy} = s^{-1} \prod_{w\in O_1(x,y)} sR_w \prod_{v\in O_2(x,y)} sR_v, \quad (26)$$

where $s = \frac{P(\overline{e_{xy}})}{P(e_{xy})} = \frac{M^F - M}{M}$. For a certain real-life network, the value of s is determined. Taking the logarithm of Eq. 26,

we can obtain

$$r'_{xy} = \underbrace{(|O_1(x,y)| + |O_2(x,y)|)logs}_{the\ number\ of\ two\ predictors} + \underbrace{\sum_{w \in O_1(x,y)} logR_w}_{the\ role\ function\ of\ the\ first\ predictor} + \underbrace{\sum_{v \in O_2(x,y)} logR_v}_{the\ role\ function\ of\ the\ second\ predictor} . \quad (27)$$

For the two-motif naive Bayes model, r'_{xy} refers to the feature score between nodes x and y. It consists of three parts, the first part refers to the sum of the number of edge-dependent motifs obtained from two motifs, the second part is the role functions of the first motif, and the third part represents the role functions of the second motif.

B. PERFORMANCE EVALUATION

In order to compare the principle difference between single-motif and two-motif, we choose a small toy network with seven nodes, as shown in Fig. 6(a). The predicted link is between nodes A and B, and the predictors containing the predicted edge (A, B) are one S_1, two S_2, two S_5 and one S_6. Combining any two predictors, there are six combinations. To simplify our analysis, we choose three of them to illustrate the experiment results. Considering that the number of nodes in a motif might have different impact on link prediction, the three combinations we selected include the combination of two 3-node predictors, one 3-node and one 4-node predictors, and two 4-node predictors as shown in Fig. 6(b). It can be seen that the two-motif method of link prediction should take into account two motif structures that contain the predicted edge simultaneously.

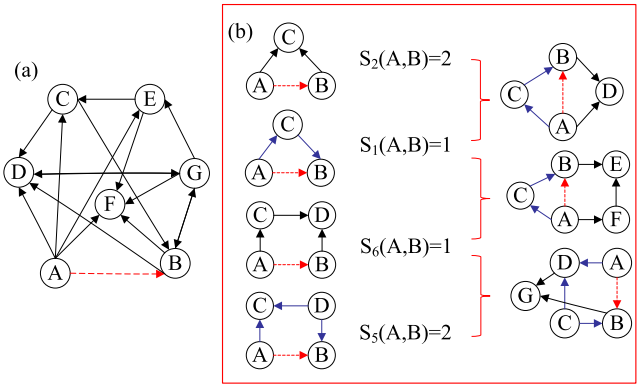

FIGURE 6. Link prediction based on two motifs. (a) A small toy network, and (b) three combination forms of two motifs.

In order to compare the performance of link prediction between single-motif and two-motif, the two-motif experiments have been performed using three combinations shown in Table 2. According to Eq. 27, the feature score of the predicted link based on two motifs is equivalent to the superposition of the feature score obtained with two single motifs. Compared to the prediction performance of the corresponding single motif, the two-motif based prediction capability can be improved in most cases.

In the two-motif naive Bayes model, the number of edge-dependent motifs is equivalent to the addition of two single motifs. If there is a strong correlation between two single motif predictors, a better prediction result can be obtained when combining the two predictors for link prediction. However, if there is no correlation between two motifs, the performance of the two-motif method may not be significantly improved, or even be degraded. Therefore, if two motifs are directly added, there is no guarantee that each combination can improve the performance of link prediction. For example, in the three combinations of FWME networks as shown in Table 2, the prediction performance of the combination of S_2 and S_5 is higher than S_2 but lower than S_5. If we directly superimpose three or more kinds of motifs together, this effect will be amplified, which means this method cannot be applied to link prediction by multiple motifs.

TABLE 2. The performance comparison of link prediction between single- and two- motif naive Bayes models.

Motif	S_1	S_2	S_5	S_6	S_1+S_2	S_2+S_5	S_5+S_6
Bison	0.210	0.419	0.306	0.387	0.371	0.306	**0.403**
FWEG	0.587	0.549	0.716	0.626	**0.607**	**0.721**	**0.738**
Macaques	0.540	0.426	0.570	0.504	0.451	**0.578**	0.561
FWME	0.594	0.522	0.708	0.597	**0.597**	0.705	**0.721**
CESF	0.679	0.511	0.761	0.642	0.665	**0.761**	**0.766**
C. elegans	0.741	0.700	0.843	0.759	**0.825**	**0.857**	**0.854**
SmaGri	0.683	0.637	0.840	0.687	**0.751**	**0.849**	**0.855**

V. LINK PREDICTION BY MULTIPLE MOTIFS USING A MACHINE LEARNING FRAMEWORK

Considering that the two-motif based method cannot be effectively extended to multiple motifs, we propose a machine learning framework for link prediction. In this study, we use the classifier XGBoost to integrate the features of multiple motifs and to obtain higher prediction performance. In order to test the predictive performance of multiple motifs, we compare it with the state-of-the-art methods and verify its effectiveness by experimental analysis. At the same time, we obtain the correlation between all the predictors by calculating Maximum Information Coefficient (MIC). The result provides a foundation for motif selection in the prediction method fusing multiple motifs, and also helps to understand the evolution mechanism of directed networks.

A. XGBOOST CLASSIFIER

XGBoost is an abbreviation for eXtreme Gradient Boosting, which is a gradient boosting machine implemented by C++. The design philosophy of XGBoost has the

TABLE 3. The prediction results of using single motif and all motifs.

Motif	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	*All*
Bison	0.376	0.435	0.331	0.363	0.371	**0.441**	0.355	0.403	0.382	0.430	0.376	0.333	**0.725**
FWEG	0.648	0.508	0.440	0.497	**0.757**	0.612	0.713	0.598	0.585	0.593	0.590	0.525	**0.868**
Macaques	0.563	0.454	0.484	0.560	0.570	0.506	0.475	0.547	**0.592**	0.506	0.534	0.511	**0.599**
FWME	0.646	0.508	0.404	0.481	**0.772**	0.604	0.672	0.468	0.615	0.502	0.469	0.539	**0.855**
CESF	0.686	0.455	0.445	0.495	**0.813**	0.675	0.737	0.488	0.694	0.522	0.515	0.541	**0.841**
C. elegans	0.783	0.713	0.737	0.584	**0.874**	0.765	0.778	0.654	0.825	0.732	0.724	0.777	**0.914**
SmaGri	0.682	0.637	0.654	0.500	**0.853**	0.689	0.676	0.600	0.646	0.609	0.614	0.641	**0.889**

following advantages: fast, portable, less code, and fault tolerance. The biggest advantage of XGBoost is that it can automatically parallelize the multi-threading of CPU and improve the algorithm performance. The difference between XGBoost and traditional GBDT is that the traditional GBDT only uses the first derivative information. XGBoost performs a second-order Taylor expansion on the loss function and adds a regular term to the objective function to obtain the optimal solution as a whole, thereby balancing the complexity of the objective function and preventing over-fitting [32].

XGBoost has been widely used in many aspects. Ren *et al*. proposed a classification method combining convolutional neural network and extreme gradient enhancement to improve the effectiveness of image classification [33]. Chen *et al*. used the gradient enhancer for disease prediction, and they presented a model of extreme gradient boosting machine for MiRNA-disease association prediction [34]. Zhong *et al*. compared three widely-used classifiers, including XGBoost, random forest and support vector machine, and found that XGBoost can achieve the best performance [35]. In order to explore the combined effects of multiple motifs, we use the XGBoost classifier for comprehensive link prediction.

B. PERFORMANCE EVALUATION

We use the feature scores of all the single-motif naive Bayes model in Fig. 2 as multi-dimensional features, and then use the XGBoost classifier for link prediction. The results of SMNB are shown in columns 2-13 of Table 3, and the bold data indicates the highest prediction performance for each experimental network. The XGBoost predicting results based on all the motif features are shown in the last column, which are always higher than any single motif based predictor.

In order to verify the effectiveness of the multi-motif based link prediction, we compare it with the state-of-the-art methods including potential theory, RA, LP, and SRW. RA is the abbreviation for Resource Allocation, which is a new similarity measure motivated by the resource allocation process taking place on networks [36]. LP is the abbreviation for Local Path, and the index is presented to estimate the likelihood of a link existence after adding the third-order path information [37]. SRW is the abbreviation for Superposed Random Walk, which can obtain better prediction with a lower computational complexity [38]. The experiment results for different training set size in the C.elegans network are shown in Fig. 7, which indicate that the performance of XGBoost method using all motifs is significantly higher than other existing methods, and the similar results can be obtained in the other six real-life networks.

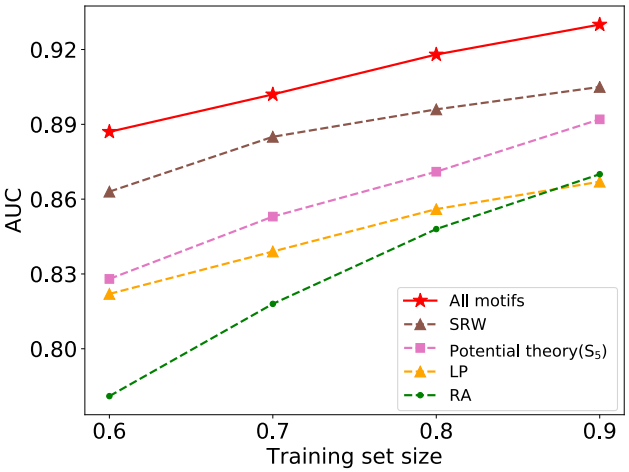

FIGURE 7. The performance comparison between all the motif characteristics and the state-of-the-art methods in the C.elegans network.

C. MOTIF CORRELATION ANALYSIS

Traditionally, the Pearson correlation coefficient is generally used to find the correlation between two variables [39]. It is a linear correlation coefficient, which is used to reflect the degree of linear correlation between two variables. The value is between 1 and -1. If two variables are completely positively correlated, the value is 1; if two variables are completely negatively correlated, the value is -1; and 0 means linearly independent. The larger the absolute value is, the stronger the linear relationship is. The Pearson correlation coefficient can be defined as

$$\rho = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{S_X}\right)\left(\frac{Y_i - \overline{Y}}{S_Y}\right), \tag{28}$$

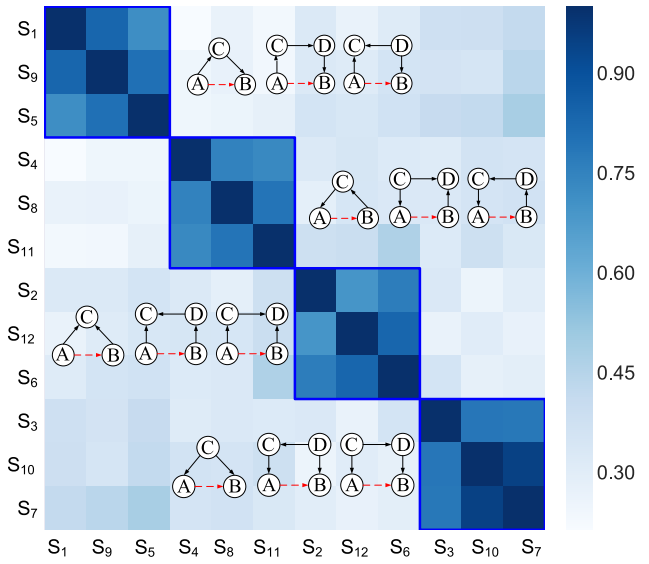

FIGURE 8. Correlation analysis among 12 kinds of motif predictors.

where $\overline{X}$ and $\overline{Y}$ are the average values of the sample values of variables X and Y, and S_X and S_Y are the standard deviations. However, the Pearson correlation coefficient cannot measure the slope of the linear and nonlinear relationship.

The statistic of Maximum Information Coefficient (MIC) [40] is superior to the traditional Pearson correlation coefficient, because it can determine the functional and non-functional relationship between variables, and obtain the variable influence in a real-life dataset. It can be defined as

$$MIC[X;Y] = \max_{|X||Y|<B} \frac{I[X;Y]}{\log_2(min(|X|,|Y|))}, \qquad (29)$$

where $I[X;Y]$ represents the mutual information between variables X and Y. $|X|$ and $|Y|$ represent the number of segments divided in the directions of X and Y respectively in the scatter plot grid, and $|X||Y|<B$ indicates that the total number of all the squares cannot be greater than B. B takes 0.6 power of the total data, which is an empirical value.

In this study, we use MIC to measure the correlation between motifs. The redundancy (i.e., correlation) between a pair of motifs f_i and f_j is defined as $MIC{=}(f_i, f_j)$. The larger the value of $MIC{=}(f_i, f_j)$ is, the stronger the redundancy between the motifs f_i and f_j is. If the value of $MIC(f_i, f_j)$ is 0, it indicates that f_i and f_j are independent of each other. The MIC correlation between all the motifs is shown in Fig. 8.

It can be seen that the stronger the correlations between the motifs are, the more similar the motif structures are, as shown in each blue box of Fig. 8. For instance, if the connected structure (an edge and two nodes) in a 4-node motif is regarded as a node, the motifs S_5 and S_9 are the same as S_1. The same result can be obtained from other boxes below. The motifs with strong correlation can be classified into one category, so the calculation for the role function of 4-node motifs is reasonable. In the two-motif naive Bayes model for link prediction, due to the randomness in the choice of motifs, there is no guarantee that the prediction performance for each combination can be improved. Analyzing the correlation between all motifs provides a way of motif feature selection for link prediction based on multiple motifs, and reduces the instability of the performance caused by random selection.

VI. CONCLUSION

In summary, we proposed a link prediction method for directed networks based on multiple motif information. Firstly, a naive Bayes model based single-motif method was proposed for link prediction and the experimental results suggest this method is superior to calculating the number of edge-dependent motifs. Next, we constructed a two-motif naive Bayes model and a multi-motif based prediction method using a framework of machine learning. The results indicate that the method fusing all the motifs can improve the performance of prediction, which is better than the state-of-the-art methods. At the same time, we used Maximum Information Coefficients (MIC) to analyze the correlation of all the predictors. The correlation between the motif predictors verified the rationality of calculating the role function of 4-node motifs by analyzing the predictor structure of the same type. Our research is helpful to understand the evolutionary mechanism of directed networks and can be extended to other types of complex networks for link prediction.

REFERENCES

[1] R. R. Sarukkai, "Link prediction and path analysis using Markov chains," *Comput. Netw.*, vol. 33, no. 1, pp. 377–386, 2000.
[2] J. Zhu, J. Hong, and J. G. Hughes, "Using Markov chains for structural link prediction in adaptive Web sites," in *User Modeling* (Lecture Notes in Computer Science), vol. 2311. Berlin, Germany: Springer, 2004, pp. 60–73.
[3] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Stat. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.
[4] L. Getoor and C. Diehl, "Link mining: A survey," *ACM SIGKDD Explorations Newslett.*, vol. 7, no. 2, pp. 3–12, 2005.
[5] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási, "Human mobility, social ties, and link prediction," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, no. 9, 2011, pp. 1100–1108.
[6] L. Zhu, D. Guo, J. Yin, G. Ver Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2765–2777, Oct. 2016.
[7] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang, "Deep belief network-based approaches for link prediction in signed social networks," *Entropy*, vol. 17, no. 4, pp. 2140–2169, 2015.
[8] M. Bilgic, G. M. Namata, and L. Getoor, "Combining collective classification and link prediction," in *Proc. 7th IEEE Int. Conf. Data Mining Workshops*, Oct. 2007, pp. 381–386.
[9] L. Lü, "Link prediction on complex networks," *J. Univ. Electron. Sci. Technol. China*, vol. 39, no. 5, pp. 651–661, 2010.
[10] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.
[11] X. Zhang, J. Liu, J. Li, and L. Liu, "Large-scale dynamic social network directed graph K-in & out-degree anonymity algorithm for protecting community structure," *IEEE Access*, vol. 7, pp. 108371–108383, 2019.
[12] P. V. Paulau, C. Feenders, and B. Blasius, "Motif analysis in directed ordered networks and applications to food Webs," *Sci. Rep.*, vol. 5, no. 1, 2015, Art. no. 11926.
[13] N. Zhao, F. Hu, Z. Li, and Y. Gao, "Simultaneous wireless information and power transfer strategies in relaying network with direct link to maximize throughput," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8514–8524, Sep. 2018.

[14] M. J. Brzozowski and D. M. Romero, "Who should i follow? Recommending people in directed social networks," in *Proc. Int. Conf. Weblogs Social Media*, 2011.
[15] Z. Jiao, H. Wang, K. Ma, L. Zou, and J. Xiang, "Directed connectivity of brain default networks in resting state using GCA and motif," *Frontiers Biosci.*, vol. 22, no. 10, pp. 1634–1643, 2017.
[16] R. Milo, S. Shen-Orr, S. 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.
[17] 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, pp. 2–9, 2013.
[18] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: A local naive Bayes model," *EPL*, vol. 96, no. 4, 2011, Art. no. 48007.
[19] J. H. Wu, G. J. Zhang, Y. Ren, and X. Y. Zhang, "Weighted local naive Bayes link prediction," *J. Inf. Process. Syst.*, vol. 13, no. 4, pp. 914–927, 2017.
[20] *Bison Network Dataset–KONECT*, 2016. [Online]. Available: http://konect.uni-koblenz.de/networks/moreno_bison
[21] D. F. Lott, "Dominance relations and breeding rate in mature male American bison," *Ethology*, vol. 49, no. 4, pp. 418–432, 1979.
[22] J. Kunegis, "KONECT: The Koblenz network collection," in *Proc. 22nd Int. Conf. World Wide Web Companion*, no. 8, 2013, pp. 1343–1350.
[23] R. Ulanowicz, J. Heymans, and M. Egnotovich, "Network analysis of trophic dynamics in South Florida ecosystems, FY 99: The graminoid ecosystem," Annu. Rep. United States Geol. Service Biol. Resour. Division, Chesapeake Biological Lab., Univ. Maryland, College Park, MD, USA, Tech. Rep. CBL 00-0176, 2000.
[24] *Macaques Network Dataset–KONECT*, 2016. [Online]. Available: http://konect.uni-koblenz.de/networks/moreno_mac
[25] Y. Takahata, "Diachronic changes in the dominance relations of adult female japanese monkeys of the arashiyama B troop," Monkeys Arashiyama, State Univ. New York Press, Albany, NY, USA, Tech. Rep., 1991, pp. 123–139.
[26] 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 Shelf Sci.*, vol. 47, no. 3, pp. 329–349, 1998.
[27] *Florida Ecosystem Dry Network dataset–KONECT*, 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/foodweb-baydry
[28] D. J. Watts and S. H. Strogatz, "Collective dynamics of small world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
[29] V. Batagelj and M. A. Pajek. *Datasets Website*. Accessed: Jan. 14, 2018. [Online]. Available: http://vlado.fmf.uni-lj.si/pub/networks/data/
[30] K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon, "Exploiting longer cycles for link prediction in signed networks," in *Proc. Int. Conf. Inf. Knowl. Manage.*, no. 6, 2011, pp. 1157–1162.
[31] A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabási, "The topological relationship between the large-scale attributes and local interaction patterns of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 52, pp. 17940–17945, 2004.
[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, no. 10, 2016, pp. 785–794.
[33] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, "A novel image classification method with CNN-XGBoost model," in *Digital Forensics and Watermarking*, vol. 10431. Berlin, Germany: Springer, 2017, pp. 378–390.
[34] X. Chen, L. Huang, D. Xie, and Q. Zhao, "EGBMMDA: Extreme gradient boosting machine for MiRNA-disease association prediction," *Cell Death Disease*, vol. 9, no. 1, p. 3, 2018.
[35] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sens. Environ.*, vol. 221, pp. 430–443, Feb. 2019.
[36] 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, 2009.
[37] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 4, 2009, Art. no. 046122.
[38] W. Liu and L. Lü, "Link prediction based on local random walk," *Europhys. Lett.*, vol. 89, no. 5, 2010, Art. no. 58007.
[39] M. Mudelsee, "Estimating pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series," *Math. Geol.*, vol. 35, no. 6, pp. 651–665, 2003.
[40] W. H. Hsu, "Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning," *Inf. Sci.*, vol. 163, nos. 1–3, pp. 103–122, Jun. 2004.

YAFANG LIU received the bachelor's degree from the School of Information and Electrical Engineering, Heibei University of Engineering, China, in 2017. She is currently pursuing the Ph.D. degree with the College of Information and Communication Engineering, Dalian Minzu University, Dalian, China. Her research interests include social network analysis, data visualization, and link prediction.

TING LI received the M.S. degree in signal processing and the Ph.D. degree in signal processing from the Dalian University of Technology (DUT), in 2005 and 2015, respectively. She is currently an Associate Professor with Dalian Minzu University. Her research interests include non-Gaussian signal processing, biomedical signal processing, and big data processing.

XIAOKE XU (M'13) received the Ph.D. degree from the College of Information and Communication Engineering, Dalian Maritime University, in 2008. He was a Postdoctoral Fellow with The Hong Kong Polytechnic University and a Visiting Scholar with the City University of Hong Kong. He is currently a Professor with the College of Information and Communication Engineering, Dalian Minzu University. His current research interests are in information spreading on complex networks, network community detection, and data mining in social networks.