Link Prediction in Signed Social Networks: From Status Theory to Motif Families

Si-Yuan Liu, Jing Xiao, and Xiao-Ke Xu, *Member, IEEE*

***Abstract*—Link prediction can discover missing information and evolution mechanism of complex networks, so a huge number of novel algorithms have been proposed recently. However, the existing link prediction algorithms for directed signed networks only depend on motifs that satisfy status theory, and other types of motifs are rarely taken into account. In this study, first we propose a link prediction method based on the number of edge-dependent motifs, and explain it by a naive Bayes model. Furthermore, we put forward a Signed Local Naive Bayes (SLNB) model based on two kinds of different motifs, which has higher prediction performance than only considering a single motif. Finally, we combine all the 3-node motifs to form a motif family, and use a machine learning framework for link prediction. The results show that motif families can greatly improve the performance of link prediction. Moreover, according to the correlation between these predictors, the intrinsic relationship between different motifs can be discovered, and the computational complexity of link prediction can be reduced after feature selection. Our research can not only improve the performance of link prediction, but also be helpful to uncover the evolutionary mechanism of signed social networks.**

***Index Terms*—Signed network, Status theory, Motif families, Naive Bayses model, Link prediction.**

I. Introduction

LINK prediction refers to predicting the existence of a link between two nodes based on the known nodes and edges in complex networks [1], [2], which is of great significance in both static and dynamic networks. For static networks, link prediction can find false and missing links [3]. For dynamic networks, it is helpful to understand the evolution mechanism of a real-life network and compare the advantages and shortcomings of different network evolution models [4]–[6]. In recent years, more and more attention has been paid to this field [7]–[9].

In all the methods of link prediction, the approaches based on structure information are the most commonly used, which can be classified into global and local structure methods [10], [11]. Although the algorithms based on global structures can achieve high performance, the computational complexity of them is relatively high and they cannot be applied to large-scale networks [12], [13]. In contrast, the prediction algorithms based on local structures have low complexity and are easy to be implemented [3], [10], [14]. However, most of them can not be applied to signed networks directly, because these local methods seldom consider the sign of each edge between a pair of nodes.

Signed networks are a special type of complex networks with both positive and negative edges [15]. The positive edges represent positive relationships such as “friends” and “trust”, and are represented by the positive sign “+”. The negative edges represent negative relationships such as “enemies” and “distrust”, and are represented by the negative sign “–”. In a signed network, whether two nodes do link each other depends on not only the number of common neighbors between them, but also the sign and direction of each edge in their neighborhood.

The most commonly used link prediction algorithm in signed networks is based on small subgraphs that satisfy status theory, and these subgraphs can be understood as special cases of motifs [16]. Compared with many global structural features such as small-world [17] and scale-free [18], motif (i.e., subgraph) is the most basic structural and functional unit in complex networks [19]. Link prediction via a motif can be expressed as: whether two nodes do connect depends on the specific functional units formed by the edge connecting these two nodes and their neighbor nodes [20]. The motif-based prediction algorithm considers the connection patterns (including the signs and directions of edges) between node pairs and their neighbors, so it is applicable to signed networks [21], [22].

The existing motif-based link prediction algorithms for signed networks have the following three drawbacks. First, the current methods only focus on motifs that satisfy status theory [23]–[25], but do not consider other types of motifs. Actually, the mechanism by which the motifs can be employed to link prediction in signed networks is not explained. At the same time, there is no answer as to explain the mechanism that calculating the number of motifs on the predicted edge can be used for link prediction. Finally, the classical algorithms of link prediction are based on only a single motif and do not think about the relation between different kinds of motifs.

To solve the above mentioned problem, we investigate a novel framework based on the edge-dependent motifs for link prediction. In this study, we first use motif theory to explore the relationship between the number of each motif and its ability

Manuscript received July 10, 2019; revised October 31, 2019; accepted October 31, 2019. Date of publication November 6, 2019; date of current version September 2, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61603073 and Grant 61773091, in part by the Key Research and Development Plan of Liaoning Province (2018104016), in part by the Liaoning Revitalization Talents Program (XLYC1807106), and in part by the Program for the Outstanding Innovative Talents of Higher Learning Institutions of Liaoning (LR2016070). Recommended for acceptance by J. C. S. Lui. (*Corresponding author: Xiao-Ke Xu.*)

The authors are with the College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China (e-mail: liusiyuan0311@foxmail.com; hrbeuxiaojing@aliyun.com; xuxiaoke@foxmail.com).

Digital Object Identifier 10.1109/TNSE.2019.2951806

for link prediction. Experiments on five empirical signed networks demonstrate that the prediction ability of a motif depends not on its number in the whole network but on the number of edge-dependent motifs. Then we explain the edge-dependent motif based link prediction by a naive Bayes model. Secondly, we put forward a Signed Local Naive Bayes (SLNB) model, which has higher prediction performance than a single motif. Finally, we combine all the types of 3-node motifs to build a machine learning classifier based on motif families. The network structure information used by motif families in link prediction is more comprehensive than status theory and thus gives more accurate prediction.

The paper is organized as follows. The empirical network data and the evaluation indicators of link prediction are introduced in Section II. In Section III, we evaluate the limitation of status and motif theory for link prediction and propose a prediction algorithm based on the number of edge-dependent motifs. Moreover, the mechanism of the proposed prediction method is explained by a naive Bayes model. A Signed Local Naive Bayes (SLNB) consisting of two motifs is proposed and a machine learning predictor of motif families is built to improve the performance of link prediction in Section IV. We finally offer the conclusion in Section V.

II. Data Description and Evaluation Indicators

A. Description of Empirical Network Data

In this study, we use five signed social networks where links are explicitly positive or negative for link prediction: **Bitcoinalpha**, **Bitcoinotc**, **Wiki-RfA**, **Slashdot** and **Epinions**.

Bitcoinalpha and **Bitcoinotc** are two who-trusts-whom networks of people who trade using bitcoin on the platform called Bitcoin Alpha and the platform called Bitcoin OTC, respectively [26], [27]. Users of these platforms rate other users in a scale of -10 (total distrust) to +10 (total trust) in steps of 1. The sign of a user's rating for another user is the sign of the directed edge formed by the two users, and the absolute value of the rating indicates the weight of the edge. The Bitcoinalpha network has 3,783 nodes and 24,186 edges, where the positive edges account for 93.65% of all edges, and the negative edges account for 6.35%. The Bitcoinotc network contains 5,881 nodes and 35,592 edges, where the positive edges account for 89.99% of all edges, and the negative edges account for 10.01%.

Wiki-RfA is a network of voting between Wikipedia members [28]. To make a Wikipedia editor to be an administrator, a candidate or another community member must submit a Request for Adminship (RfA). Subsequently, any Wikipedia member can vote for support, neutrality or opposition. This induces a directed, signed network in which nodes represent Wikipedia members, positive and negative edges represent supporting and opposing votes, respectively. This network contains 10,835 nodes and 185,626 edges, the ratios of positive and negative edges are 77.82% and 22.18%, respectively.

Slashdot is a network formed by users tagging each other as friends or foes on the website called Slashdot which is a technology-related news website known for its specific user community [29]. This network was obtained in February 2009 which nodes represent users and friend/foes tags represent positive/negative edges. This network contains 82,144 nodes and 549,202 edges, and the ratios of positive and negative edges are 77.4% and 22.6%, respectively.

Epinions is a who-trust-whom online social network of a general consumer review site Epinions.com [29]. Members of the site can decide whether to "trust" each other. The "trust" relationship between two members represents a positive edge. Conversely, the "distrust" relationship between two members represents a negative edge. This network contains 131,828 nodes and 841,372 edges, and the ratios of positive and negative edges are 85.3% and 14.7%, respectively.

It should be noted that we only use the direction and sign information of edges, regardless of the weights of edges in the above five empirical signed networks.

B. Evaluation Indicators for Link Prediction

We use the two most commonly used evaluation indicators, AUC [30] and Precision [31], to measure the performance of link prediction. AUC refers to the area under the Receiver Operating Characteristic (ROC) curve. The networks used throughout this study are large-scale, so we do not draw the ROC curve when actually calculating the value of AUC, but utilize the sample comparison method to obtain an approximate value.

AUC can be understood as the probability that the score of a randomly chosen missing edge is higher than the score of a randomly chosen non-existent edge. That is to say, each time randomly selects a missing edge and a non-existent edge. If the score of the missing edge is higher than the score of the non-existent edge, add 1 point; if the two scores are equal, add 0.5 points; if the score of the missing edge is lower than the score of the non-existent edge, no points will be added. After performing this process N times independently, there are X times when the score of the missing edge is higher than that of the non-existent edge, and Y times when the scores of the two edges are equal. Then AUC can be defined as

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

AUC measures the overall performance of link prediction [30]. Precision does not focus on the overall performance, but only pay attention to whether a few edges with high scores are accurately predicted [31]. Precision is defined as the proportion of edges that are accurately predicted in the set of the top L scores. After sorting the scores of edges in descending order, if there are m missing edges in the set of top L scores, then Precision can be defined as

$$Precision = \frac{m}{L}. \tag{2}$$

The value of Precision is related to the parameter L. When the parameter is given, the larger the value of Precision is, the more accurate the Prediction is.

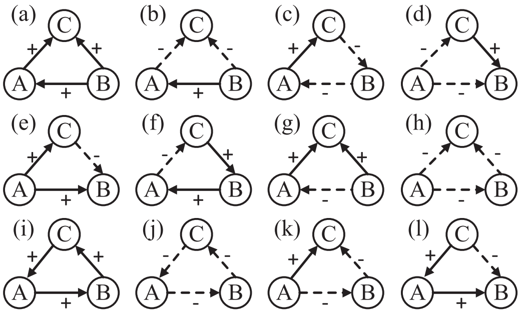

Fig. 1. All the signed loop-embedded motifs of order 3 with single directional edges. (a)-(h) Motifs satisfy status theory; (i)-(l) Motifs do not satisfy status theory.

III. From Status Theory to Edge-Dependent Motif Theory

A. Status Theory in Signed Networks

Signed networks are a special type of social networks that have both positive and negative edges. Status theory of signed networks was first proposed by Davis [32]. The theory holds that the sign of an edge connecting two nodes depends on the status difference between them. The detailed description is as follows. A positive link from A to B means that B has a higher status than A. In contrast, a negative link from A to B means that B has a lower status than A. These relative levels of status can be propagated along multi-step paths of signed links [29]. For example, if there is a positive link from B to A and a positive link from A to C in Fig. 1(a), according to the transitivity of relative status, the status of C is higher than B.

Based on the definition of status theory, it is possible to find all the motifs of different orders that satisfy or do not satisfy the theory. The clustering mechanism makes a network more inclined to form loops [21]. Moreover, high order motifs and low order motifs mutually define and predict each other [33]. Therefore, we only consider loop-embedded motifs of order 3. Eight motifs that satisfy status theory and four motifs that do not satisfy the theory are presented in Fig. 1. The former are shown in Figs. 1(a)-(h) and the latter are illustrated in Figs. 1(i)-(l). The solid black line connecting two nodes indicates a positive link and the black dotted line implies a negative link.

Taking Fig. 1(i) as an example, the reason why the motif does not satisfy status theory is as follows. The positive link from C to A indicates that the status of A is higher than C, and the positive link from B to C indicates that the status of C is higher than B. According to the transitivity of the relative level of status, the status of A is higher than B. However, there is a positive link from A to B in Fig. 1(i) (i.e., the status of B is higher than A), and this contradicts the conclusion derived from the transitivity of relative status. Therefore, this motif does not satisfy status theory.

B. Limitation of Status Theory for Link Prediction

As the basic unit of networks, motif can uncover the structural design principles of networks to a certain extent [16]. A network satisfies status theory if there are many motifs that satisfy status theory in the network, and these high-frequency motifs in the network perform well for link prediction [10], [34]. Therefore, status theory can be evaluated by link prediction [21]. In this section, we evaluate the limitation of status theory for link prediction by the experimental results.

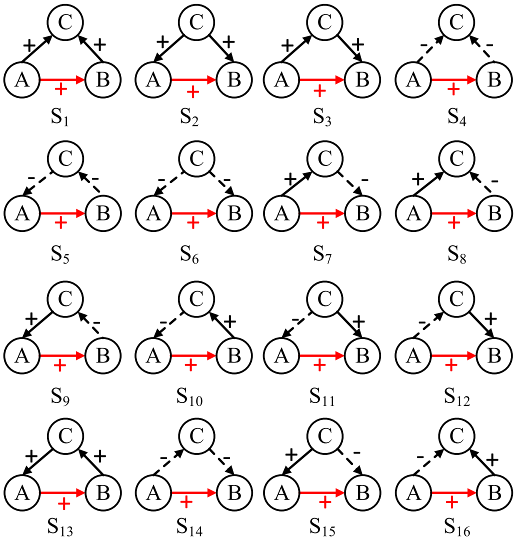

Fig. 2. The predictors for positive edges corresponding to the motifs in Fig. 1.

There are three predicted edges on each motif in Fig. 1, so 12 motifs correspond to 36 predictors in total. After removing the repeated predictors, there are a total of 32 predictors. Since there are both positive and negative edges in signed networks, we divide the missing edges into positive and negative missing edges, and then perform link prediction for these two types of edges respectively. Therefore, 32 predictors can be divided into 16 predictors for positive edges and 16 predictors for negative edges, as shown in Figs. 2 and 3, respectively. In these two figures, S_1-S_{12} and N_1-N_{12} represent predictors that satisfy status theory, and S_{13}-S_{16} and N_{13}-N_{16} represent predictors that do not satisfy status theory. The black solid lines indicate the positive edges, the black dashed lines indicate the negative edges, the red solid lines and the red dashed lines indicate positive and negative predicted edges (i.e., the missing edges and the non-existent edges which are introduced in Section II-B), respectively.

We use 16 predictors for positive edges in Fig. 2 and the same number of predictors for negative edges in Fig. 3 for link prediction, respectively. The result of the predictors for positive edges is shown in Fig. 4. The circular scatter points represent the predictors that satisfy status theory and the triangular scatter points represent the predictors that do not satisfy the theory. It can be seen that most predictors that satisfy status theory have high prediction performance, which indicates that these empirical networks are basically in line with status theory. However, not all the predictors that satisfy the theory have higher prediction performance than the predictors that do not satisfy the theory. In addition, it can also be found that some predictors that do not satisfy status theory can achieve very high prediction performance.

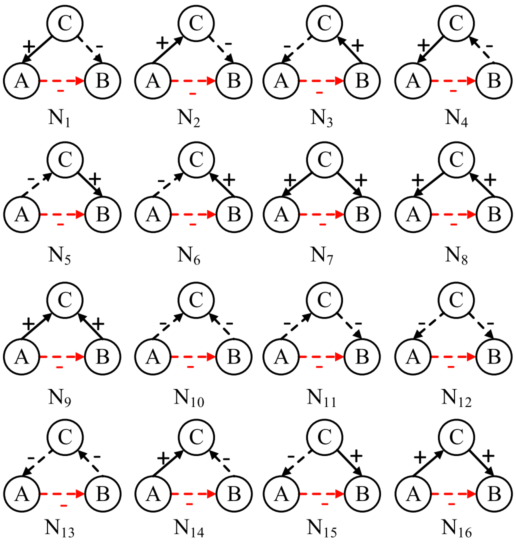

Fig. 3. The predictors for negative edges corresponding to the motifs in Fig. 1.

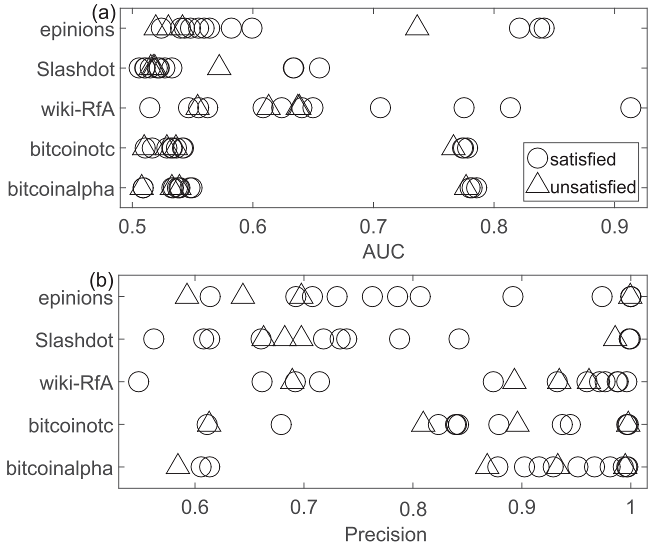

Fig. 4. The results of evaluating status theory using the predictors for positive edges. "Satisfied" means the result of the predictors that satisfy status theory and "Unsatisfied" is the result of the predictors that do not satisfy status theory. (a) The prediction result of AUC, and (b) the prediction performance measured by Precision.

In order to demonstrate the limitation of status theory more specifically, we compare the average performance of all predictors that satisfy status theory and that of all predictors that do not satisfy the theory, as shown in Fig. 5. In a statistical sense, the predictors that satisfy status theory do not outperform the predictors that do not satisfy this theory. The same result can be obtained when using AUC and precision metrics to measure the predictive performance. Therefore, status theory is not very effective for link prediction. Actually, link prediction using the predictors for negative edges can also lead to such a conclusion.

The result of evaluating status theory by link prediction indicates that the theory is not sufficient for link prediction. As a subset of motif theory, status theory only considers the motifs that satisfy status theory. Therefore, we try to take into account the motifs that do not satisfy status theory, and attempt to use motif theory for link prediction in singed networks.

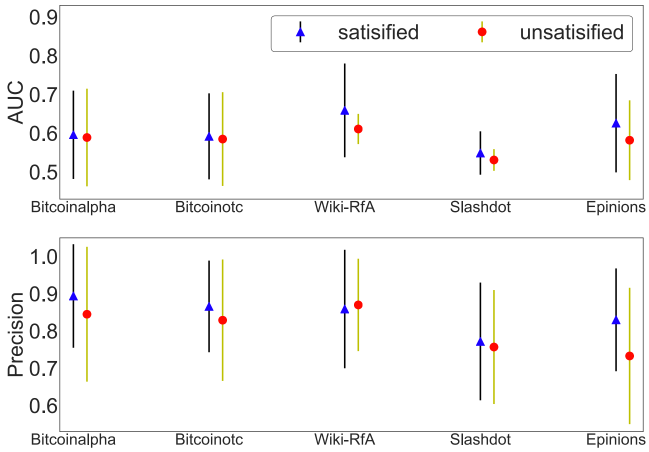

Fig. 5. The comparison of the average performance of all predictors that satisfy status theory and that of all predictors that do not satisfy the theory. "Satisfied" means the result of the predictors that satisfy status theory and "Unsatisfied" is the result of the predictors that do not satisfy status theory. (a) The prediction result of AUC, and (b) the prediction performance measured by precision.

C. Limitation of Motif Theory for Link Prediction

According to the similar mechanism of status theory, the idea of the traditional motif theory for link prediction is as follows. The more times a certain motif appears in a real-life network, the more important the functional module corresponding to the motif, which means that the number of motifs is very important. Put this idea on link prediction, it can be considered that the more times a certain motif appears in the network, the stronger the predictive ability of the corresponding predictor of the motif. Next, we examine whether this idea is valid by studying the relationship between the number of motifs and the performance of link prediction.

We calculate the number of motifs corresponding to each predictor in five empirical networks. The relationship between the number of each motif and the performance of the predictor corresponding to the motif is shown in Fig. 6. We find that except the AUC values of Slashdot and Epinions and the Precision value of Slashdot, there is a weak correlation between the number of motifs and the performance of link prediction. To prove this weak correlation, we calculate the Pearson correlation coefficient [35], [36] between the number of motifs and the performance of link prediction. The Pearson correlation coefficients between the number of motifs and the AUC values of Bitcoinalpha, Bitcoinotc, Wiki-RfA, Slashdot and Epinions are 0.611, −0.251, 0.398, 0.867, 0.828, respectively. The Pearson correlation coefficients between the number of motifs and the Precision values of these five networks are 0.478, 0.271, 0.415, 0.763, 0.572, respectively.

Regardless of whether AUC or Precision is used as an evaluation indicator, the above results demonstrate that the ability of a motif for link prediction depends weakly on the total number of this motif in a real-life network. There are two main

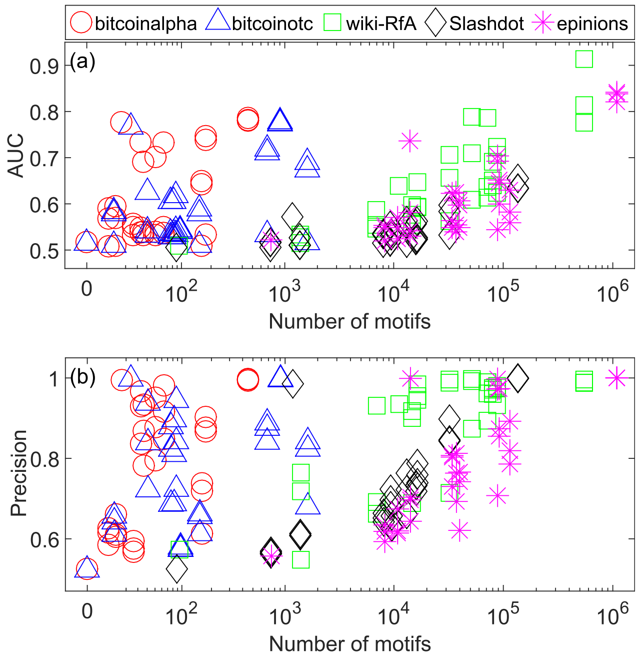

Fig. 6. The relationship between the number of each motif and its corresponding prediction performance in real-life networks.

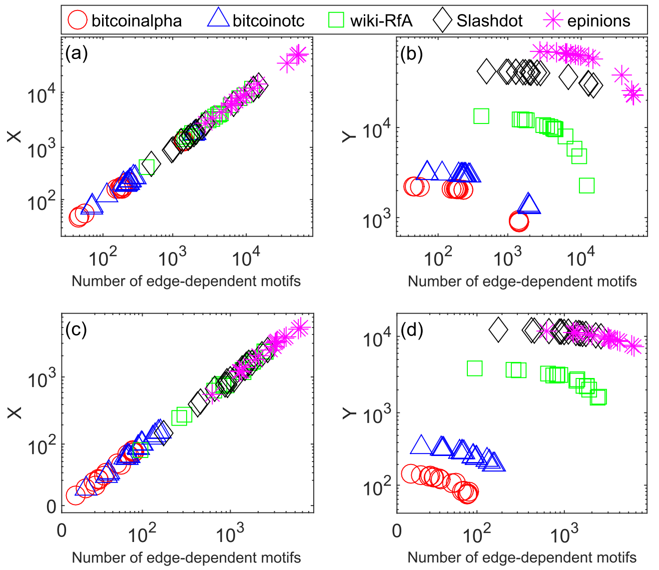

Fig. 7. The relationship between the number of edge-dependent motifs corresponding to a predictor and the value of X (and Y) when calculating AUC. (a)-(b) results of the predictors for positive edges; (c)-(d) results of the predictors for negative edges.

reasons for this. First, although some motifs have a very large number in empirical networks, the distribution of these motifs in the network is heterogeneous. That is to say, a certain motif is present in a large amount at some local structures, but such type of motif is very sparse in other parts. In addition, the highly abundant motifs cannot exist in isolation but aggregate into a large motif cluster [33], and this causes the motifs to be concentrated on a small number of edges. It should be noted that the strong correlation between the total number of each motif and the prediction performance of Slashdot and Epinions may be due to the fact that the motifs in these networks are evenly distributed on each predicted edge.

Moreover, we also find that the performances of distinct predictors corresponding to the same motif are different. For example, the predictors S_4, N_1 and N_2 correspond to the same motif, but for the Bitcoinalpha network, the AUC results of the three predictors are 0.534, 0.739, and 0.747, respectively. And the Precision results are 0.902, 0.876, and 0.867, respectively. This also demonstrates that the predictive performance of the motif is dependent on the specific predicted edge. The conclusion also applies to the other four real-life networks.

D. Link Prediction by Edge-Dependent Motif Theory

In the above section, it has been observed that there is no strong correlation between the number of motifs and the performance of link prediction. This is primarily due to the non-uniformity and agglomeration of the motif distribution in real-life networks. Therefore, we conjecture that the predictive ability of each motif does not depend on the total number of it, but on the number of edge-dependent motifs. For a certain motif, the edge-dependent motif refers to the motif that exists on the predicted edge. That is to say, the edge-dependent motifs corresponding to a certain motif in the network are a subset of all such motifs. The idea of the edge-dependent motif can be expressed as follows. For an edge, the more number of one motif it involves, the more prominent the functional module corresponding to the motif is. Put this idea on link prediction, the greater the probability that such a motif appears on each edge is, the stronger the link prediction ability of the motif is.

Based on the above analysis, we attempt to explain how the number of edge-dependent motifs affects AUC in this section. We first explore the relationship between the number of edge-dependent motifs and the value of X (and Y) when calculating AUC in Section II-B. The number of edge-dependent motifs is proportional to the value of X, and inversely proportional to the value of Y in Fig. 7. This shows that when calculating the value of AUC, the more the number of edge-dependent motifs, the more the situations where the score of a missing edge is greater than that of a non-existent edge, and the less the situations where the two scores are equal.

Next, we explore the relationship between the number of edge-dependent motifs and AUC. The results are shown in Figs. 8(a) and (c), there is a strong correlation between the number of edge-dependent motifs and AUC. Figures 8(b) and (d) show the relationship between the number of edge-dependent motifs and Precision, and there is also a strong correlation between them. Then, we calculate the Pearson correlation coefficient between the number of edge-dependent motifs and the link prediction performance to prove the strong correlation between them. When using the predictors for positive edges for link prediction, except that the Pearson correlation coefficient between the number of edge-dependent motifs and AUC in the Bitcoinalpha network is 0.999, the coefficients in other four networks are all 1.000. The Pearson correlation coefficients between the number of edge-dependent motifs and Precision in Bitcoinalpha, Bitcoinotc, Wiki-RfA, Slashdot and Epinions are 0.560, 0.711, 0.773, 0.900 and 0.853, respectively. The same

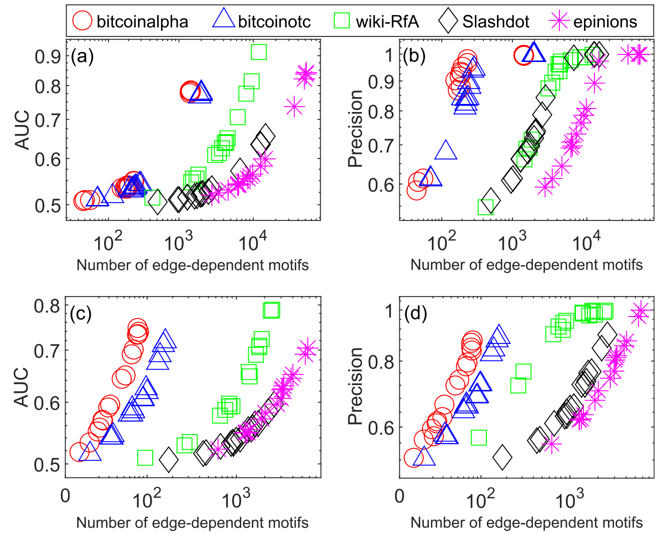

Fig. 8. The relationship between the number of edge-dependent motifs corresponding to a predictor and the performance of the motif predictor. (a)-(b) results of the predictors for positive edges; (c)-(d) results of the predictors for negative edges.

results can be obtained when the predictors for negative edges are used, which indicates that the correlation between the link prediction performance and the number of edge-dependent motifs is stronger than that between the link prediction performance and the number of motifs. The same conclusion can be achieved when using AUC and precision metrics to measure the predictive performance. The above results support our conjecture that the ability of a motif for link prediction depends on not its number in the whole network but the number of edge-dependent motifs. Our findings suggest that the greater the probability that such a motif appears on each edge, the stronger the link prediction ability of the motif.

E. Naive Bayes Explanation for Edge-Dependent Motif Theory

The traditional common-neighbor-based link prediction algorithms are based on the assumption that each common neighbor contributes equally to a potential link. However, Zhang *et al.* argued that this assumption was not realistic and they proposed a Local Naive Bayes (LNB) model as a probabilistic model for link prediction [37]. The key idea is that they introduced a node role function to quantify the contribution of each common neighbor to the potential link. However, the LNB model has a disadvantage of not considering the weight information of each edge. Consequently, Wu *et al.* extended the LNB model to weighted networks to overcome this shortcoming, and proposed a Weighted Local Naive Bayes (WLNB) model [38].

LNB only considers the simplest triads without weights, directions and signs, and WLNB adds the function of weighted triads. However, both LNB and WLNB are only utilized in unsigned networks. In other words, the negative edges are ignored or treated as positive edges. Thus, we propose a Signed Local Naive Bayes (SLNB) model for link prediction in signed networks. Our model considers the directed signed triads. Therefore, our model of the local subgraph (i.e., motif) has a different meaning from LNB and WLNB.

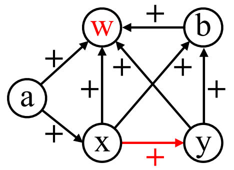

Fig. 9. The schematic diagram of calculating the score of a predicted edge (i.e., (x, y)) based on the naive Bayes model.

For a given signed network $G(V, E)$ and a train set partition E^T, suppose x and y are two nodes in the network, we use e_{xy} to indicate that the two nodes are connected, and $\overline{e_{xy}}$ to indicate that x and y are not connected. The posterior probability that nodes x and y are connected to each other or not can be calculated as:

$$P(e_{xy}|S_i(x,y)) = \frac{P(e_{xy}) \cdot P(S_i(x,y)|e_{xy})}{P(S_i(x,y))}, \tag{3}$$

$$P(\overline{e_{xy}}|S_i(x,y)) = \frac{P(\overline{e_{xy}}) \cdot P(S_i(x,y)|\overline{e_{xy}})}{P(S_i(x,y))}, \tag{4}$$

where $S_i(x, y)$ represents a set of nodes that form predictor S_i with x and y. Taking predictor S_1 as an example, there are two nodes (i.e., w and b) can form predictor S_1 with x and y in Fig. 9. Therefore, $S_1(x, y)$ is the set of nodes $\{w, b\}$.

At this time, for a given pair of nodes, we can judge whether they are more inclined to link by comparing the probability that they do connect $P(e_{xy}|S_i(x,y))$ and the probability that they do not connect $P(\overline{e_{xy}}|S_i(x,y))$. To better compare which edges are more likely to occur, we can use the ratio of the two probabilities to calculate a score for each pair of nodes:

$$\begin{aligned} r_{xy} &= \frac{P(e_{xy})}{P(\overline{e_{xy}})} \cdot \frac{P(S_i(x,y)|e_{xy})}{P(S_i(x,y)|\overline{e_{xy}})} \\ &= \frac{P(e_{xy})}{P(\overline{e_{xy}})} \cdot \prod_{w \in S_i(x,y)} \frac{P(w|e_{xy})}{P(w|\overline{e_{xy}})} \\ &= \frac{P(e_{xy})}{P(\overline{e_{xy}})} \cdot \prod_{w \in S_i(x,y)} \frac{P(\overline{e_{xy}})}{P(e_{xy})} \cdot \frac{P(e_{xy}|w)}{P(\overline{e_{xy}}|w)}. \end{aligned} \tag{5}$$

Here, $P(e_{xy})$ represents the prior probability that nodes x and y are connected, and $P(\overline{e_{xy}})$ represents the prior probability that nodes x and y are unconnected. $P(e_{xy})$ and $P(\overline{e_{xy}})$ can be calculated as:

$$P(e_{xy}) = \frac{M}{M^F}, \tag{6}$$

$$P(\overline{e_{xy}}) = \frac{M^F - M}{M^F}, \tag{7}$$

where $M^F = 2|V|(|V| - 1)$ indicates the number of all possible edges in the signed network, and $M = |E^T|$ indicates the number of existent edges in the network. $P(e_{xy}|w)$ denotes the probability of interconnection between a pair of nodes that form predictor S_i with node w. It can be expressed as:

$$P(e_{xy}|w) = \frac{N_{\triangle S_{iw}}}{N_{\triangle S_{iw}} + N_{\wedge S_{iw}}}, \tag{8}$$

where $N_{\triangle S_{iw}}$ and $N_{\wedge S_{iw}}$ represent the number of interconnected pairs of nodes and the number of pairs of nodes that are not connected to each other in the neighborhood that predictor S_i with node w, respectively. Still taking predictor S_1 as an example to analyze the role of node w in the network, nodes a, b, x and y in Fig. 9 can form this predictor with node w. Among all node pairs composed of these nodes, (a, b) and (a, y) are two unconnected edges, and (a, x), (b, x), (b, y) and (x, y) are four connected edges. Consequently, the values of $N_{\wedge S_{iw}}$ and $N_{\triangle S_{iw}}$ are 2 and 4, respectively. Therefore, it is possible to obtain a probability that a pair of neighbors that form predictor S_i with node w are not connected to each other:

$$P(\overline{e_{xy}}|w) = 1 - P(e_{xy}|w) = \frac{N_{\wedge S_{iw}}}{N_{\triangle S_{iw}} + N_{\wedge S_{iw}}}. \tag{9}$$

Combining Equations (6)-(9), r_{xy} can be simplified as:

$$r_{xy} = n^{-1} \prod_{w \in S_i(x,y)} n \frac{N_{\triangle S_{iw}} + 1}{N_{\wedge S_{iw}} + 1}, \tag{10}$$

where $n = \frac{P(\overline{e_{xy}})}{P(e_{xy})} = \frac{M^F - M}{M}$. In order to prevent the expression from being meaningless because the denominator is 0, the numerator and denominator in the equation are added 1. At this point, given a node w, its role function based on predictor S_i can be defined as:

$$R_{S_{iw}} = \frac{N_{\triangle S_{iw}} + 1}{N_{\wedge S_{iw}} + 1}. \tag{11}$$

Thus, Equation (10) can be described as:

$$r_{xy} = n^{-1} \prod_{w \in S_i(x,y)} n R_{S_{iw}}. \tag{12}$$

Obviously, for a given network, n is a constant. Take the logarithm of Equation (12):

$$r'_{xy} = log(n^{-1} \prod_{w \in S_i(x,y)} n R_{S_{iw}}) = \sum_{w \in S_i(x,y)} log(n R_{S_{iw}})$$
$$= \underbrace{|S_i(x,y)| logn}_{r_1} + \underbrace{\sum_{w \in S_i(x,y)} log R_{S_{iw}}}_{r_2}. \tag{13}$$

If $R_{S_{iw}} = 1$ for each node, r'_{xy} degenerates into the product of the number of edge-dependent motifs introduced in Section III-D and a constant, and the SLNB model is also degraded to the edge-dependent-motif based link prediction algorithm.

There are two parts when calculating the scores of edges by the SLNB model. The first part r_1 can be measured by the number of predictor S_i composed of a predicted edge and their common neighbors. The second part r_2 focuses on the contribution of different common neighbors to the potential link, which is the sum of the role function of each common neighbor. The node role function is of high computational complexity, so we explore the influence of the second part on link prediction. If this part has little effect on the prediction result, it is feasible to omit this part for link prediction.

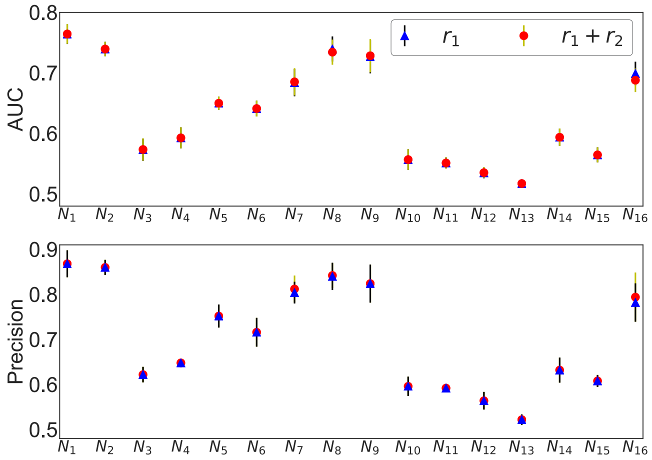

Fig. 10. The performance comparison of link prediction using the predictors for negative edges based on the entire SLNB model (i.e., $r_1 + r_2$) and those based on only part r_1 of the model.

We test the performance of the predictors for positive and negative edges based on r_1 and $r_1 + r_2$ respectively, and the results of negative predictors are shown in Fig. 10. The difference between the results of using the entire SLNB model (i.e., $r_1 + r_2$) and the results of the first part r_1 is very small. In these five directed signed empirical networks, our results suggest that the role function has little effect on the performance of link prediction. The same conclusion can be obtained when using AUC and precision metrics to measure the predictive performance. Therefore, it implies that the second part of the naive Bayes model can be negligible for link prediction, and the same conclusion can be obtained using the predictors for positive edges.

IV. From Single Motif to Motif Families

A. *Signed Local Naive Bayes Model of Two Motifs*

Different motifs in the same network can aggregate around hubs [33]. This phenomenon may lead to two situations. First, for the dense region of edges, multiple motifs are widespread on these edges. For example, a pair of nodes can form distinct predictors with different nodes, such as nodes A and B in Fig. 11 can form predictor S_1 with node C_1, and can also form predictor S_4 with node C_2. In addition, for the sparse region of edges, the number of any motif may be very few. It is difficult to get a desired result by a single motif predictor, and a better prediction result may be achieved by integrating the information of two motifs. Therefore, we attempt to propose a SLNB model consisting of two motifs for link prediction in this section.

When we use predictors S_i and S_j for link prediction, the posterior probability that nodes x and y are connected to each other or not can be calculated as:

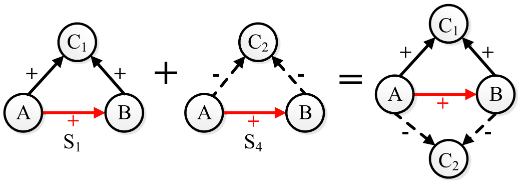

Fig. 11. Link prediction based on two kinds of motifs.

$$\begin{aligned} &P(e_{xy}|(S_i(x,y),S_j(x,y))) \\ &= \frac{P(e_{xy})\cdot P((S_i(x,y),S_j(x,y))|e_{xy})}{P(S_i(x,y),S_j(x,y))} \\ &= \frac{P(e_{xy})\cdot P(S_i(x,y)|e_{xy})\cdot P(S_j(x,y)|e_{xy})}{P(S_i(x,y),S_j(x,y))}, \end{aligned} \quad (14)$$

$$\begin{aligned} &P(\overline{e_{xy}}|(S_i(x,y),S_j(x,y))) \\ &= \frac{P(\overline{e_{xy}})\cdot P((S_i(x,y),S_j(x,y))|\overline{e_{xy}})}{P(S_i(x,y),S_j(x,y))} \\ &= \frac{P(\overline{e_{xy}})\cdot P(S_i(x,y)|\overline{e_{xy}})\cdot P(S_j(x,y)|\overline{e_{xy}})}{P(S_i(x,y),S_j(x,y))}, \end{aligned} \quad (15)$$

where $S_i(x,y)$ and $S_j(x,y)$ represent two sets of nodes that form predictors S_i and S_j with a pair of nodes (x,y) respectively. At this time, the score of this node pair can be calculated as:

$$\begin{aligned} r_{xy} &= \frac{P(e_{xy}|(S_i(x,y),S_j(x,y)))}{P(\overline{e_{xy}}|(S_i(x,y),S_j(x,y)))} \\ &= \frac{P(e_{xy})}{P(\overline{e_{xy}})}\cdot\frac{P(S_i(x,y)|e_{xy})}{P(S_i(x,y)|\overline{e_{xy}})}\cdot\frac{P(S_j(x,y)|e_{xy})}{P(S_j(x,y)|\overline{e_{xy}})}, \end{aligned} \quad (16)$$

where $\frac{P(S_i(x,y)|e_{xy})}{P(S_i(x,y)|\overline{e_{xy}})}$ can be calculated as:

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

Similarly, $\frac{P(S_j(x,y)|e_{xy})}{P(S_j(x,y)|\overline{e_{xy}})}$ can be calculated as:

$$\frac{P(S_j(x,y)|e_{xy})}{P(S_j(x,y)|\overline{e_{xy}})} = \prod_{v\in S_j(x,y)}\frac{P(\overline{e_{xy}})}{P(e_{xy})}\cdot\frac{P(e_{xy}|v)}{P(\overline{e_{xy}}|v)}. \quad (18)$$

Combining Equations (16)-(18), r_{xy} can be simplified as:

$$\begin{aligned} r_{xy} = &\frac{P(e_{xy})}{P(\overline{e_{xy}})}\cdot\underbrace{\prod_{w\in S_i(x,y)}\frac{P(\overline{e_{xy}})}{P(e_{xy})}\cdot\frac{P(e_{xy}|w)}{P(\overline{e_{xy}}|w)}}_{the\ influence\ of\ predictor\ S_i} \\ &\cdot\underbrace{\prod_{v\in S_j(x,y)}\frac{P(\overline{e_{xy}})}{P(e_{xy})}\cdot\frac{P(e_{xy}|v)}{P(\overline{e_{xy}}|v)}}_{the\ influence\ of\ predictor\ S_j}. \end{aligned} \quad (19)$$

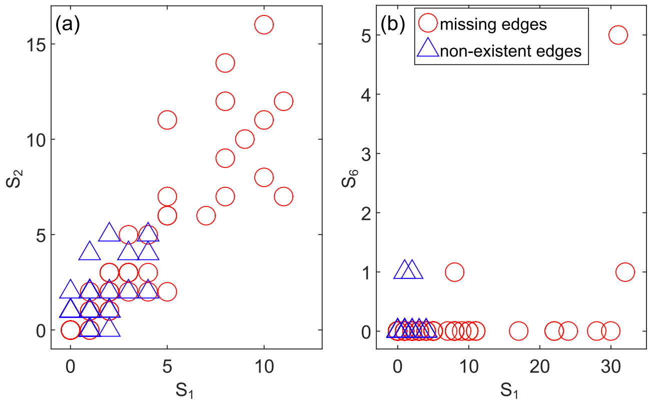

Fig. 12. The correlation detection between different predictors. (a) The correlation result between predictors S_1 and S_2, and (b) the correlation result between predictors S_1 and S_6.

According to the process of Equations (5)-(12), Equation (19) can be simplified as:

$$r_{xy} = n^{-1}\prod_{w\in S_i(x,y)} nR_{S_{i\,w}}\prod_{v\in S_j(x,y)} nR_{S_{j\,v}}. \quad (20)$$

Here, $n = \frac{P(\overline{e_{xy}})}{P(e_{xy})} = \frac{M^F-M}{M}$, which is introduced in Section III-E. $R_{S_{i\,w}} = \frac{N_{\triangle S_{i\,w}}+1}{N_{\wedge S_{i\,w}}+1}$ is the role function of node w based on predictor S_i, and $R_{S_{j\,v}} = \frac{N_{\triangle S_{j_v}}+1}{N_{\wedge S_{j_v}}+1}$ is the role function of node v based on predictor S_j. We use a logarithmic function on both sides of Equation (20), it can be written as

$$\begin{aligned} r'_{xy} = &(|S_i(x,y)|+|S_j(x,y)|)logn \\ &+\underbrace{\sum_{w\in S_i(x,y)} logR_{S_{i\,w}}}_{\substack{the\ role\ function\\ of\ predictor\ S_i}}+\underbrace{\sum_{v\in S_j(x,y)} logR_{S_{j_v}}}_{\substack{the\ role\ function\\ of\ predictor\ S_j}}. \end{aligned} \quad (21)$$

In Section III-E, we verified that it is feasible to predict links using only the first part of the SLNB model. Therefore, when performing link prediction based on two motifs, the score of the pair of nodes can be expressed as:

$$r'_{xy} = (|S_i(x,y)|+|S_j(x,y)|)logn, \quad (22)$$

where $logn$ is a constant in a real-life network.

In general, link prediction based on two motifs has better performance than that based on only a single motif. But the results based on two motifs are not always stable. Moreover, there are about 120 combinations for 16 predictors, because any two predictors can generate a combination, which is difficult to select the suitable predictors. Actually, it is difficult to determine the relationship between any two predictors and sometimes they are correlated. For example, there is a strong correlation between predictors S_1 and S_2, while predictors S_1 and S_6 are irrelevant, as shown in Fig. 12. Therefore, a better prediction result might be obtained when combining predictors

TABLE I
THE RESULTS OF LINK PREDICTION BY COMBINING MULTIPLE POSITIVE PREDICTORS. HERE P REPRESENTS THE RESULT OF PRECISION

Motif	Bitcoinalpha		Bitcoinotc		Wiki-RfA		Slashdot		Epinions	
	AUC	P	AUC	P	AUC	P	AUC	P	AUC	P
S_1	0.782	**0.996**	0.775	0.997	0.814	0.988	0.634	0.999	0.838	**1.000**
S_2	0.780	0.993	0.774	0.996	0.775	0.988	0.634	0.998	0.821	**1.000**
S_3	**0.786**	0.996	**0.778**	**0.997**	**0.913**	**0.996**	**0.655**	**1.000**	**0.841**	**1.000**
S_4	0.534	0.902	0.533	0.840	0.608	0.874	0.515	0.661	0.543	0.708
S_5	0.509	0.605	0.511	0.611	0.515	0.548	0.506	0.562	0.524	0.614
S_6	0.509	0.613	0.517	0.679	0.563	0.714	0.533	0.843	0.599	0.974
S_7	0.533	0.878	0.529	0.823	0.650	0.972	0.527	0.788	0.582	0.892
S_8	0.539	0.951	0.537	0.879	0.641	0.960	0.523	0.739	0.560	0.786
S_9	0.548	0.966	0.542	0.937	0.555	0.692	0.510	0.608	0.548	0.731
S_{10}	0.539	0.929	0.533	0.839	0.547	0.662	0.511	0.614	0.564	0.807
S_{11}	0.537	0.916	0.533	0.842	0.624	0.933	0.521	0.718	0.555	0.763
S_{12}	0.550	0.981	0.543	0.945	0.706	0.976	0.522	0.733	0.540	0.692
S_{13}	0.777	0.995	0.766	0.998	0.638	0.935	0.572	0.985	0.736	0.999
S_{14}	0.508	0.585	0.510	0.613	0.554	0.689	0.515	0.663	0.519	0.593
S_{15}	0.533	0.868	0.529	0.810	0.613	0.893	0.518	0.697	0.541	0.697
S_{16}	0.539	0.933	0.536	0.896	0.637	0.962	0.517	0.682	0.530	0.644
All	**0.823**	**0.996**	**0.825**	**0.998**	**0.959**	**0.998**	**0.746**	**1.000**	**0.899**	**1.000**

TABLE II
THE RESULTS OF LINK PREDICTION BY COMBINING MULTIPLE NEGATIVE PREDICTORS. HERE P REPRESENTS THE RESULT OF PRECISION.

Motif	Bitcoinalpha		Bitcoinotc		Wiki-RfA		Slashdot		Epinions	
	AUC	P	AUC	P	AUC	P	AUC	P	AUC	P
N_1	0.739	**0.876**	**0.718**	**0.888**	0.710	0.992	0.554	0.723	0.692	0.972
N_2	**0.747**	0.867	0.710	0.876	**0.789**	**0.997**	0.564	0.764	**0.704**	**0.998**
N_3	0.569	0.610	0.578	0.642	0.528	0.719	0.517	0.570	0.571	0.674
N_4	0.595	0.661	0.584	0.657	0.533	0.764	0.516	0.566	0.548	0.619
N_5	0.649	0.738	0.690	0.842	0.706	0.996	**0.597**	**0.897**	0.644	0.855
N_6	0.643	0.720	0.674	0.823	0.656	0.989	0.584	0.846	0.652	0.874
N_7	0.702	0.795	0.605	0.686	0.691	0.975	0.563	0.760	0.629	0.819
N_8	0.734	0.837	0.625	0.722	0.586	0.931	0.526	0.609	0.623	0.803
N_9	0.732	0.849	0.619	0.724	0.724	0.981	0.557	0.732	0.625	0.812
N_{10}	0.550	0.595	0.540	0.580	0.594	0.954	0.531	0.629	0.549	0.621
N_{11}	0.556	0.576	0.544	0.574	0.647	0.984	0.538	0.657	0.597	0.743
N_{12}	0.533	0.568	0.540	0.574	0.592	0.946	0.531	0.630	0.607	0.766
N_{13}	0.518	0.525	0.515	0.523	0.509	0.573	0.506	0.525	0.522	0.557
N_{14}	0.591	0.626	0.589	0.662	0.597	0.917	0.534	0.642	0.548	0.619
N_{15}	0.569	0.615	0.579	0.656	0.577	0.900	0.536	0.652	0.554	0.634
N_{16}	0.691	0.782	0.606	0.688	0.786	0.990	0.553	0.723	0.581	0.703
All	**0.886**	**0.990**	**0.876**	**1.000**	**0.936**	**0.998**	**0.729**	**1.000**	**0.829**	**1.000**

S_1 and S_6 for link prediction, while the performance may be not significantly improved when combining S_1 and S_2.

B. Link Prediction by Motif Families

In addition to the unsupervised prediction method described in the previous sections, we can also combine multiple predictors to perform link prediction through a machine-learning classifier. We use two ways to construct motif families in our research when using a machine learning classifier for prediction: one way is to combine all the predictors for positive edges or all the predictors for negative edges, and the other way is to combine the predictors after feature selection. In this study, we use the XGBoost Classifier as a machine learning model.

1) XGBoost Classifier: XGBoost (i.e., eXtreme Gradient Boosting) is a gradient boosting machine implemented by C++. Different from the traditional GBDT model which 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 find the overall optimal solution, which is used to weigh the complexity of the objective function and the model, thus it can prevent overfitting [39]. In addition to the theoretical differences with the GBDT model, XGBoost also has the following advantages in terms of actual performance: fast, scalable, less code-writing, and fault-tolerant.

2) Link Prediction by All Motifs: We combine all the predictors for positive edges and all the predictors for negative edges to construct motif families, respectively. Treat the scores of edges calculated by 16 predictors as 16-dimensional features, and then use XGboost for link prediction. The prediction results of the five large-scale signed networks based on all the predictors for positive and negative edges are shown in Tables I and II, respectively. In each column, the best result and the result based on the motif family (all the motifs) are highlighted in boldface. From these two tables, link prediction using the motif family is more accurate than using a single motif, and this conclusion can be drawn from all the five experimental networks. The motif families not only consider the motifs that satisfy status theory, but also utilize the motifs that do not satisfy status theory, so they have the highest prediction performance.

Then, we compare our proposed method (i.e., motif family) with two state-of-the-art methods in signed networks: FriendTNS [25], [40] and status theory [23], [24]. The results of the predictors for positive edges in the Bitcoinalpha network are shown in Fig. 13, as the size of the training set increases, the predictive performance of all methods is improved. Furthermore, motif-based methods (i.e., motif family and status theory) can obtain higher prediction results than FriendTNS, and the method of motif family obtains the best predictive performance because it considers more types of motifs (i.e., motifs that do not satisfy status theory) than status theory. It can also be found that as the size of the training set increases, the predictive performance of status theory gradually approaches that of motif family, which shows that our method has obvious superiority when the training size is small, and the superiority decreases as the training size increases. In order to demonstrate the performance of our proposed method on all datasets, we compare it with the existing algorithms in five real-world signed networks when the training size is 0.6. The results are shown in Fig. 14, the method of motif family has the best predictive performance in all the networks.

3) Motif Correlation Detection and Feature Selection: We use the Pearson correlation coefficient to characterize the correlation between two features, and calculate the Pearson correlation matrix [35], [36] between all features by the scores of

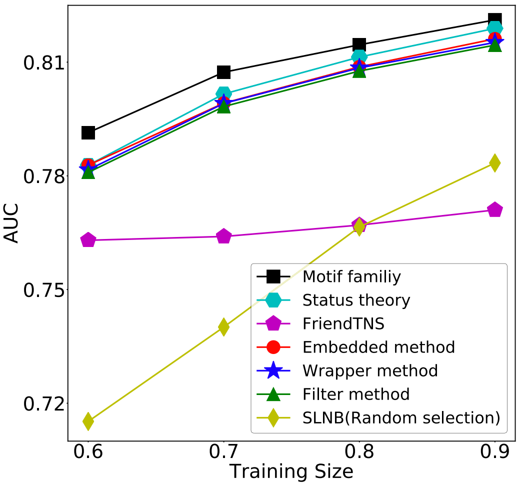

Fig. 13. The comparison of our proposed method with the existing methods in the Bitcoinalpha network.

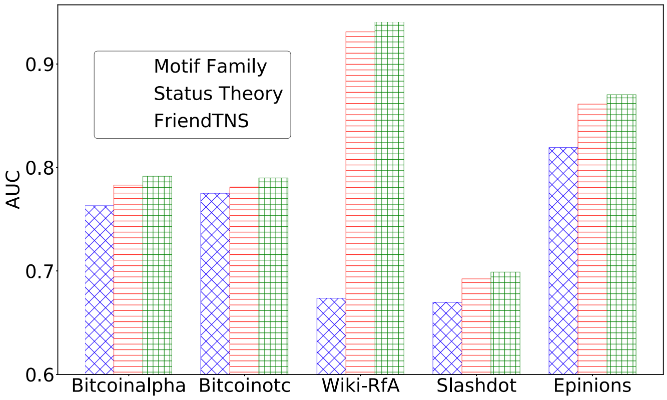

Fig. 14. The comparison of our proposed method with the existing methods in five real signed networks.

edges calculated by all the predictors for positive edges and all the predictors for negative edges. The Pearson correlation matrix of the predictors for positive and negative edges are shown in Figs. 15 and 16 respectively. Both the predictors for positive and negative edges can be divided into nine categories, and all predictors in each black box are of a class. For example, four types of motifs (i.e., S_1, S_2, S_3, S_{13}) are formed by two positive edges and the predicted edge, and there is a strong correlation between any two of the four motifs. In fact, these four motifs belong to the same type of features in link prediction. Conversely, another four types of motifs (i.e., S_4, S_5, S_6, S_{14}) which are formed by two negative edges and the predicted edge are independent features, and the correlation among them is very weak. Mining the correlation between the motifs helps to understand the evolution mechanism of signed networks [41], [42].

In order to reduce the computational complexity, we use four methods for feature selection when using a motif family for link prediction. These four feature selection methods are embedded method [43], wrapper method [44], filter method [45] and the SLNB model of two motifs in Section IV-A, respectively. As shown in Fig. 13, the link prediction after feature selection by the SLNB model of two motifs has the worst performance. This is because when using this model for link prediction, the two predictors are randomly selected. The link prediction results after feature selection by the other three methods are very close to that of using all the predictors. The results show that link prediction based on a motif family after feature selection can reduce the computational complexity while maintaining relatively high performance.

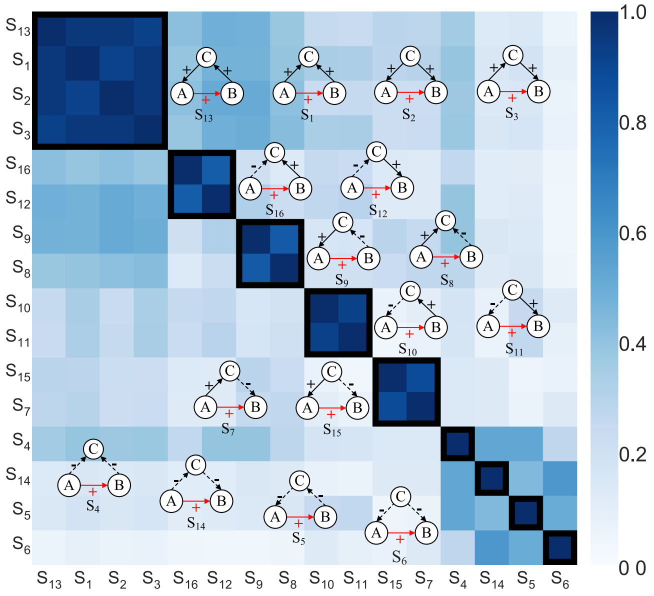

Fig. 15. The Pearson correlation matrix of 16 predictors for positive edges.

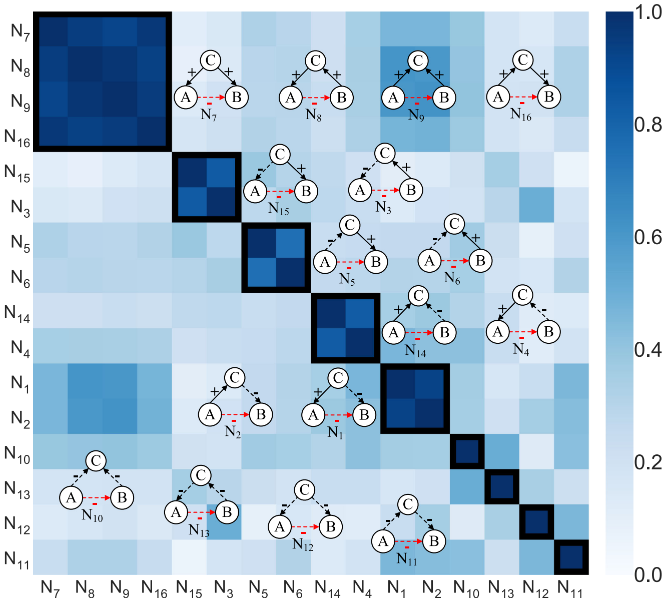

Fig. 16. The Pearson correlation matrix of 16 predictors for negative edges.

V. Conclusion

In summary, we investigate a comprehensive motif-based framework for link prediction in signed networks. In this study, we proposed a novel link prediction algorithm based on the number of edge-dependent motifs and explained it by a naive Bayes model. Furthermore, we proposed a Signed Local Naive Bayes (SLNB) model consisting of two motifs, which has higher prediction performance than the model based on only a single motif. Finally, we combine all the 3-node motifs to form a motif family, and use a machine learning classifier for link prediction. One approach of link prediction based on the motif family is to combine all the predictors for positive edges or all the predictors for negative edges. Another way is to combine the motifs after feature selection. The results show that link prediction by motif families can improve the performance of link prediction. Moreover, link prediction after feature selection can reduce the computational complexity on the premise of lower prediction performance reduction.

Our research can not only improve the performance of link prediction, but also be helpful to uncover the correlation relationship between different motifs and the evolutionary mechanism of signed networks. In future research, we will expand our framework from link prediction to sign and weight prediction in signed networks [26], [34], [46], [47]. Recently, various graph neural network methods [48]–[50] have been proposed to improve the performance of link prediction. These methods mainly use network embedding and deep graph neural networks to automatically learn arbitrary motif features instead of using predefined motifs in this study. Although it is not suitable to directly apply these deep learning methods to signed networks, a promising direction is combining and comparing our proposed methods with them in future.

References

[1] A. Popescul and L. H. Ungar, "Statistical relational learning for link prediction," in *Proc. Int. Joint Conf. Artif. Intell. Workshop Learn. Statist. Models Relational Data*, 2003, pp. 81–90.

[2] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 659–666.

[3] 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.

[4] W. Wang, Q. Zhang, and T. Zhou, "Evaluating network models: A likelihood analysis," *Europhys. Lett.*, vol. 98, no. 2, pp. 28004–28009, 2012.

[5] B. Kaya and M. Poyraz, "Supervised link prediction in symptom networks with evolving case," *Measurement*, vol. 56, no. 6, pp. 231–238, 2014.

[6] M. Xiao, J. Liao, S. M. Djouadi, and Q. Cao, "Lips: Link prediction as a service for data aggregation applications," *Ad Hoc Netw.*, vol. 19, pp. 43–58, 2014.

[7] H. Wang, W. Hu, and Z. Qiu, "Nodes' evolution diversity and link prediction in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2263–2274, Oct. 2017.

[8] A. De, S. Bhattacharya, S. Sarkar, N. Ganguly, and S. Chakrabarti, "Discriminative link prediction using local, community, and global signals," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2057–2070, Aug. 2016.

[9] Z. Wang, J. Liang, R. Li, and Y. Qian, "An approach to cold-start link prediction: Establishing connections between non-topological and topological information," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 2857–2870, Nov. 2016.

[10] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A Statist. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[11] Z. Ren, A. Zeng, and Y. Zhang, "Structure-oriented prediction in complex networks," *Phys. Rep.*, vol. 750, pp. 1–51, 2018.

[12] 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.

[13] R. Guimerá and M. Salespardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 52, pp. 22 073–22 078, 2009.

[14] T. Zhou, L. Lü, and Y. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, 2009.

[15] D. Cartwright and F. Harary, "Structural balance: A generalization of Heider's theory," *Psychological Rev.*, vol. 63, pp. 277–293, 1956.

[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] S. Goyal, M. V. D. Leij, and J. L. Moraga-Gonzlez, "Economics: An emerging small world," *J. Political Economy*, vol. 114, no. 2, pp. 403–412, 2006.

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

[19] X. Xu, J. Zhang, and S. Michael, "Superfamily phenomena and motifs of networks induced from time series," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 50, pp. 19601–19605, 2008.

[20] F. Aghabozorgi and M. R. Khayyambashi, "A new similarity measure for link prediction based on local structures in social networks," *Physica A: Statist. Mech. Appl.*, vol. 501, pp. 12–23, 2018.

[21] Q. Zhang, L. Lü, W. Wang, and T. Zhou, "Potential theory for directed networks," *Plos One*, vol. 8, no. 2, 2013, Art. no. e55437.

[22] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proc. Nat. Acad. Sci. USA*, vol. 115, pp. E11221–E11230, 2018.

[23] X. Li, "Towards practical link prediction approaches in signed social networks," in *Proc. 26th Conf. User Model., Adapt. Personalization*, 2018, pp. 269–272.

[24] S. Gu, L. Chen, B. Li, W. Liu, and B. Chen, "Link prediction on signed social networks based on latent space mapping," *Appl. Intell.*, vol. 49, no. 2, pp. 703–722, 2019.

[25] P. Symeonidis and E. Tiakas, "Transitive node similarity: Predicting and recommending links in signed social networks," *World Wide Web*, vol. 17, no. 4, pp. 743–776, 2014.

[26] S. Kumar, F. Spezzano, V. Subrahmanian, and C. Faloutsos, "Edge weight prediction in weighted signed networks," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 221–230.

[27] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "REV2: Fraudulent user prediction in rating platforms," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 333–341.

[28] R. West, H. S. Paskov, J. Leskovec, and C. Potts, "Exploiting social network structure for person-to-person sentiment analysis," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 297–310, 2014. [Online]. Available: http://arxiv.org/abs/1409.2450

[29] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2010, pp. 1361–1370.

[30] 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.

[31] J. B. Schafer, F. Dan, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.

[32] J. A. Davis, "Clustering and structural balance in graphs," *Social Netw.*, vol. 20, no. 2, pp. 27–33, 1977.

[33] A. Vázquez, 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.

[34] A. Papaoikonomou, M. Kardara, K. Tserpes, and T. A. Varvarigou, "Predicting edge signs in social networks using frequent subgraph discovery," *IEEE Internet Comput.*, vol. 18, no. 5, pp. 36–43, Sep./Oct. 2014.

[35] J. Gravier, V. Vignal, S. Bissey-Breton, and J. Farre, "The use of linear regression methods and Pearsons correlation matrix to identify mechanical–physical–chemical parameters controlling the micro-electrochemical behaviour of machined copper," *Corrosion Sci.*, vol. 50, no. 10, pp. 2885–2894, 2008.

[36] A. M. Gadermann, M. Guhn, and B. D. Zumbo, "Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide," *Practical Assessment Res. Eval.*, vol. 17, no. 1, pp. 1–13, 2012.

[37] Z. Liu, Q. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: A local nave Bayes model," *Europhys. Lett.*, vol. 96, 2011, Art. no. 48007.

[38] J. Wu, G. Zhang, Y. Ren, X. Zhang, and Q. Yang, "Weighted local naive Bayes link prediction," *J. Inf. Process. Syst.*, vol. 13, no. 4, pp. 914–927, 2017.

[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[40] P. Symeonidis, E. Tiakas, and Y. Manolopoulos, "Transitive node similarity for link prediction in social networks with positive and negative links," in *Proc. ACM Conf. Recommender Syst.*, 2010, pp. 183–190.

[41] J. Tang, H. Gao, and H. Liu, "mTrust: Discerning multi-faceted trust in a connected world," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 93–102.

[42] J. Tang, H. Gao, H. Liu, and A. Das Sarma, "eTrust: Understanding trust evolution in an online world," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 253–261.

[43] M. You, J. Liu, G. Li, and Y. Chen, "Embedded feature selection for multi-label classification of music emotions," *Int. J. Comput. Intell. Syst.*, vol. 5, no. 4, pp. 668–678, 2012.

[44] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Inf. Sci.*, vol. 179, no. 13, pp. 2208–2217, 2009.

[45] D. Zhang, S. Chen, and Z. Zhou, "Constraint score: A new filter method for feature selection with pairwise constraints," *Pattern Recognit.*, vol. 41, no. 5, pp. 1440–1451, 2008.

[46] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 641–650.

[47] T. Dubois, J. Golbeck, and A. Srinivasan, "Predicting trust and distrust in social networks," in *Proc. IEEE 3rd Int. Conf. Privacy*, 2012, pp. 418–424.

[48] M. Zhang and Y. Chen, "Weisfeiler–Lehman neural machine for link prediction," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 575–583.

[49] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–14.

[50] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5171–5181.

Si-Yuan Liu received the bachelor's degree in communication in 2017 from Dalian Minzu University, Dalian, China, where she is currently working toward the Ph.D. degree with the College of Information and Communication Engineering. Her research interests include social network analysis, network community detection, and link prediction.

Jing Xiao received the Ph.D. degree in signal and information processing from Harbin Engineering University, Harbin, China. She has done Postdoctoral training in automation from Harbin Engineering University. She is currently a Lecturer of information and communication engineering with Dalian Minzu University, Dalian, China. Her current research interests include network community detection, swarm intelligence computing, and many-objective optimization.

Xiao-Ke Xu received the Ph.D. degree from the College of Information and Communication Engineering, Dalian Maritime University, Dalian, China, in 2008. He was the 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, Dalian, China. His current research interests are in community detection, link predicting, and data mining on complex networks.