Information Sciences 541 (2020) 316–331

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Sign prediction by motif naive Bayes model in social networks

Si-Yuan Liu, Jing Xiao, Xiao-Ke Xu *

College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China

ARTICLE INFO

Article history:
Received 24 October 2019
Received in revised form 8 May 2020
Accepted 30 May 2020
Available online 7 June 2020

Keywords:
Signed network
Motif
Sign prediction
Naive Bayes model

ABSTRACT

Sign prediction is a significant research content in signed social networks, so it has attracted increasing attention from the field of online social networks recently. Traditionally, the basic idea of motif-based predictive method is to calculate the motif number on the predicted edge (i.e., the single edge-dependent motif based method), and then use a machine learning predictor for sign prediction. Although this intuition-based method can achieve great performance for sign prediction, up to now its reasonability has not been proved theoretically. Furthermore, the method of counting the number of edge-dependent motif can not distinguish the distinct role of each node on the neighborhood of the predicted edge. In this study, firstly we propose a Single Motif Naive Bayes (SMNB) model for sign prediction, which can not only explain why the single edge-dependent motif based method is efficient, but also quantify the role of each neighbor node (for 3-node predictors) or neighbor edge (for 4-node predictors) which is connected by the predicted edge for the task of sign prediction. Then, we extend SMNB by merging two types of motifs, and propose a Two Motif Naive Bayes (TMNB) model. Experimental results on real-world networks indicate that the proposed algorithms outperform the state-of-the-art approaches. Finally, we explore the intrinsic relationship among different motifs according to the matrix of Maximal Information Coefficient (MIC). Our research not only extends the traditional motif theory by proving the rationality of the edge-dependent motif based method and distinguishing a node (or an edge) contribution for sign prediction, but also is helpful to further understand the evolution mechanism of signed social networks based on the correlation among different motifs.

1. Introduction

Signed social networks are a special type of complex networks with both positive and negative edges [1]. 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 “−”. Sign prediction is a very fundamental and significant issue in signed social networks [2], which has a wide range of applications in various domains [3], such as recommendation systems, candidate elections, collaborative systems, etc. Therefore, in recent years, a large number of researchers have begun to investigate the sign inference problem in online social networks [3,4].

* Corresponding author.
E-mail address: xuxiaoke@foxmail.com (X.-K. Xu).

https://doi.org/10.1016/j.ins.2020.05.128
0020-0255/

Traditionally, the methods of sign prediction are mainly based on matrix operations [5,6]. That is to say, researchers regard a signed network as a special matrix, and then predict the sign of an edge by trust propagation, matrix factorization and matrix completion. For example, Guha et al. incorporated the distrust relationship into a propagation model for sign prediction [5]. Kunegis et al. proposed to apply spectral graph algorithms to signed networks for sign prediction [6]. With the rapid development of machine learning technology, a large number of supervised learning algorithms for sign prediction have been proposed [2,7,8], which can obtain higher performance than the algorithms based on matrix operations [9]. Among all the supervised methods, researchers mostly use network structures (including local and global structures) to predict link signs, and the local structures are the most commonly used because of their lower computational complexity [10].

As the basic structural and functional module of a real-world network [11], motif (i.e., subgraph) has been utilized in the task of link and sign prediction [12,13]. Liu et al. proposed the edge-dependent motif based method for link prediction in unsigned directed networks [14], and they explained it by a local naive Bayes model. The results show that the role function of naive Bayes models can improve the prediction performance in real-life networks. Then, we extended this method to signed networks, and proposed a Signed Local Naive Bayes (SLNB) model for link prediction [15]. The results show that the Bayes role function has no effect on link prediction, but the proposed edge-dependent motifs can greatly improve the prediction performance. It should be noted that the two works are both about link prediction, but we focus on anther problem of sign prediction in this study. Furthermore, we focus on the roles of 3-node and 4-node motifs in undirected signed networks, rather than 3-node motif in directed signed networks [15].

The idea of the single edge-dependent motif in sign prediction is to calculate the motif number on the predicted edge (i.e., an edge whose sign needs to be predicted), then use a machine learning framework to predict the edge sign. One shortcoming of this approach is that there is no theoretical support and explanation for this intuitive method. In addition, it assumes that the nodes that form a certain motif with the predicted edge have the same contribution for sign prediction. That is to say, when inferring the sign of a predicted edge, the role of each neighbor node is considered equal. However, this assumption is not always accurate and effective in a real-world network. Recently, Zhang et al. [16] and Wu et al. [17] attempted to use local naive Bayes models to distinguish the role of each node, but they only used the motifs of 3-node and did not take the advantage of high-order motifs. Moreover, these algorithms do not sufficiently fuse the information induced from multiple motifs, and the intrinsic relationship between different motifs has not been uncovered yet.

To address the aforementioned problems, in this study we firstly propose a Single Motif Naive Bayes (SMNB) model, which can explain the prediction mechanism of the single edge-dependent motif based method and quantify the role of each neighbor node (for 3-node predictors) or neighbor edge (for 4-node predictors) in the task of sign prediction. Then, we extend SMNB by merging the information of two kinds of motifs, and propose a Two Motif Naive Bayes (TMNB) model. Experimental results on real-world networks validated the effectiveness and superiority of the proposed methods comparing with the existing algorithms. Finally, the intrinsic relationship between different motifs is detected according to the matrix of the Maximal Information Coefficient (MIC) [18].

The paper is organized as follows. The empirical network data for sign prediction are introduced in Section 2. In Section 3, we introduce the problem definition and evaluation indicator of sign prediction. The Single Motif Naive Bayes (SMNB) model is proposed in Section 4. After that, we extend this model by merging any two kinds of motifs, and propose the Two Motif Naive Bayes (TMNB) model in Section 5. Finally, we conclude this paper in Section 6.

2. Data description

In this study, we test the performance of proposed algorithms on four online social networks: **Bitcoinalpha**, **Bitcoinotc**, **Wiki-RfA** and **Slashdot**.

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 [19]20. In these platforms, users 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.

Wiki-RfA is a network of voting between Wikipedia members [21]. For a Wikipedia editor to become 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.

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 [22]. This network was obtained in February 2009 which nodes represent users and friend/foes tags represent positive/negative edges.

Note that we only use the sign information of edges, regardless of the direction and weight of each edge in the above four empirical signed networks. In the process of transforming a directed signed network into an undirected signed network, both the bidirectional negative edges and the unidirectional negative edges are all treated as negative edges with no direction. Similarly, both the bidirectional positive edges and the unidirectional positive edges are all treated as positive edges. The bidirectional edges containing one positive edge and one negative edge are considered to be incompatible links, and are removed from the networks. The characteristics of these four signed networks are shown in the Table 1.

Table 1
The statical measures of four real-life signed networks. "whole" represents the original network, "positive" and "negative" represent the positive and negative subnetworks divided from the original network. n and m are the number of nodes and edges, and r is assortative coefficient [23]. t and d represent the statistics of transitivity [24] and density [25], respectively.

Measures	Bitcoinalpha			Bitcoinotc			Wiki-RfA			Slashdot		
	Positive	Negative	Whole	Positive	Negative	Whole	Positive	Negative	Whole	Positive	Negative	All
n	3783	3783	3783	5881	5881	5881	11,221	11,221	11,221	82,140	82,140	82,140
m	12,937	1187	14,124	18,574	2918	21,492	132,987	38,774	171,761	382,167	118,314	500,481
d	0.002	0.001	0.002	0.001	0.001	0.001	0.002	0.001	0.003	0.001	3.507	0.001
t	0.074	0.013	0.078	0.057	0.008	0.059	0.129	0.034	0.133	0.025	0.005	0.024
r	−0.155	−0.245	−0.169	−0.145	−0.231	-0.164	−0.057	−0.084	−0.072	−0.067	−0.167	−0.073

3. Problem definition and evaluation indicator

3.1. Problem definition

For a given undirected signed network $G(V, E, S)$, where V is the set of nodes, E is the set of edges, and S is the set of signs corresponding to the set of edges E. Let $s(A, B)$ denote the sign of edge $(A, B), s(A, B) = 1$ if the sign of edge (A, B) is positive, and $s(A, B) = -1$ if the sign of edge (A, B) is negative. Suppose that the signs of some edges in a network are missing, the problem is defined as predicting the missing signs of these edges, using the known sign and structure information of other edges. In this study, the training set and test set are obtained by randomly dividing a real-world network. In most link prediction and sign prediction tasks, the ratios of training set and test set are usually 90% and 10% [26,27], so we also use this partition ratio in this study.

3.2. Learning methodology and evaluation indicator

We first calculate the feature score of each edge in the training set and test set according to the structure of the training network, and then use a logistic regression to fit the features of training set and learn the coefficients. Finally, we use the learned coefficients to predict the signs of test set. The logistic regression can be expressed as

$$P(l = 1|\boldsymbol{v}) = \frac{1}{1 + e^{-(w_0 + \mathbf{w}^{\mathrm{T}}\boldsymbol{v})}}, \tag{1}$$

where $l \in \{0, 1\}$ is the edge sign, 1 represents positive while 0 represents negative, $\boldsymbol{v}$ is a vector of features, and $[\boldsymbol{w}; w_0]$ are the coefficients we estimate based on the training set. In this study, based on the characteristics of multiple motifs, we also use the XGBoost classifier to perform sign prediction, which has the following advantages: fast, scalable, less code-writing, and fault-tolerant [28].

Next, we use the Area Under the receiver operating Characteristic (AUC) curve to evaluate our proposed algorithms [13,29]. For a given binary classifier f and a training set $(x_i, y_i)_{i=1}^{n}$ with $x_i \in E$ and $y_i \in \{-1, 1\}$, let $P = \{x_i | y_i = 1\}$ represent the set of positive samples, and $N = \{x_i | y_i = -1\}$ represents the set of negative samples. The AUC can be defined as:

$$AUC = \frac{1}{|P||N|} \sum_{x_i \in P} \sum_{x_j \in N} I, \tag{2}$$

where I is 1 if $f(x_i) > f(x_j)$ and 0 otherwise. $|P|$ and $|N|$ are the number of positive samples and the number of negative samples, respectively. The value of AUC is essentially the probability that the random element of one set is greater than that of another set. When each positive sample is ranked higher than all the negative samples, the value of AUC is 1. For a random ranking, AUC is near 0.5.

4. Sign Prediction by Single Motif Naive Bayes model

4.1. Single Edge-dependent Motif based method

According to the clustering mechanism of complex networks, a network tends to form different sizes of loops [30]. Higher-order loops have better predictive performance than lower-order loops because they provide more information for predictive tasks, but when the order exceeds 4, the accuracy of the prediction is almost unchanged compared with 4-node loops [7]. Therefore, we only use loop-embedded motifs of orders 3 and 4 for sign prediction. All the 3-node and 4-node motifs in signed networks are shown in Fig. 1, where the solid line indicates a positive edge, and the dotted line implies a negative edge.

Any edge of each motif in Fig. 1 can be considered as a prediction edge. According to different sign combinations of edges other than this edge, we can obtain nine kinds of motif predictors, which represent those motifs that contain a prediction edge, as shown in Fig. 2. The black solid line indicates a positive edge, the black dashed line indicates a negative edge,

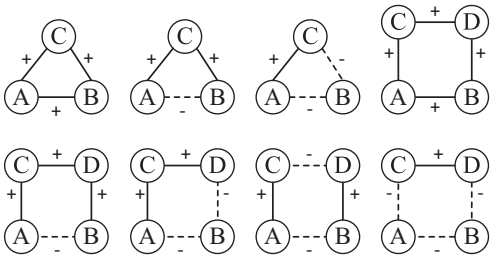

Fig. 1. All the motifs of 3-node and 4-node in signed networks.

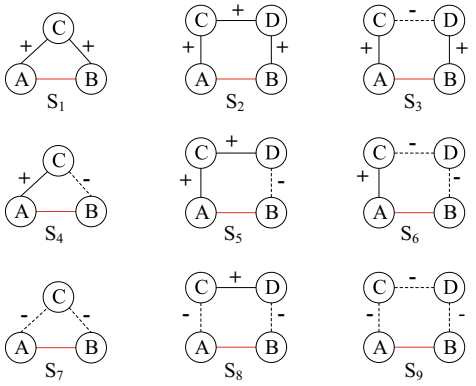

Fig. 2. The nine predictors corresponding to the motifs in Fig. 1.

and the red line indicates a predicted edge. The Single Edge-Dependent Motif based method (SEDM) can be expressed as: the score of the predicted edge is the number of certain motifs formed by the edge and other nodes in its neighborhood. For example, Leskovec et al. [2] proposed using 16 triads involving the predicted edge to predict the edge sign. Chiang et al. [7] extended their work to boost the predictive performance by considering higher-order subgraphs. Papaoikonomou et al. [31] believe that the subgraphs that emerge frequently in signed social networks possess the enough discriminative power to accurately predict edge signs, and fusing the information induced from these subgraphs can greatly improve the performance of sign prediction. Specifically, when using the predictor S_i to infer the sign of the predicted edge (A, B), the score of this edge can be calculated as

$$r_{AB} = |S_i(A, B)|, \tag{3}$$

where $S_i(A, B)$ denotes a set of nodes (for 3-node predictors) or edges (for 4-node predictors) constituting the predictor S_i with the predicted edge (A, B). $|S_i(A, B)|$ is the number of predictor S_i, i.e., the number of motifs composed of the predicted edge and other nodes. Suppose there are two nodes (i.e., M and N) in the network which can form two S_1 with the predicted edge (A, B), as shown in Fig. 3. Therefore, when using the predictor S_1 to calculate the score of this edge, $r_{AB} = |S_1(A, B)| = 2$.

For a given signed network G, the score of predicted edge (A, B) can be calculated by the Single Edge-Dependent Motif (SEDM) based Algorithm. Its calculation is shown in Algorithm 1.

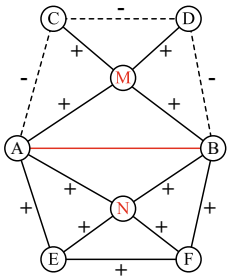

Fig. 3. The roles of different nodes forming the predictor S_1 with the predicted edge.

Algorithm 1. The Single Edge-Dependent Motif based Algorithm

Input: $G(V, E, S)$: the signed network
(A, B) : the predicted edge
S_i : the predictor for sign prediction
k : the order of S_i
Output: r'_{AB} : the score of the predicted edge (A, B)
If k equals to 3
Initialize $S_1Num = S_4Num = S_7Num = 0$
For each node M in G
If (M, A) and (M, B) are positive edges (i.e., node M can form S_1 with (A, B)) **do**
$S_1Num += 1$
⋮
Else If (M, A) and (M, B) are negative edges (i.e., node M can form S_7 with (A, B)) **do**
$S_7Num += 1$
endfor
endif
If k equals to 4
Initialize $S_2Num = S_3Num = S_5Num = S_6Num = S_8Num = S_9Num = 0$
For each edge (e_1, e_2) in G
If (e_1, e_2), (e_1, A), (e_2, B) are all positive edges or (e_1, e_2), (e_1, B), (e_2, A) are all positive edges
(i.e., edge (e_1, e_2) can form S_2 with (A, B)) **do**
$S_2Num += 1$
⋮
Else If (e_1, e_2), (e_1, A), (e_2, B) are all negative edges or (e_1, e_2), (e_1, B), (e_2, A) are all negative edges
(i.e., edge (e_1, e_2) can form S_9 with (A, B)) **do**
$S_9Num += 1$
endfor
endif
If S_i equals to S_1
$r'_{AB} = S_1Num$
⋮
Else If S_i equals to S_9
$r'_{AB} = S_9Num$
endif

The method of SEDM can be regarded as an extension of the notion of transitivity in social networks. Transitivity can be described as follows. If node A is connected to node B and node B to node C, then there is a heightened probability that node A will also be connected to node C. That is, when two nodes have a common neighbor node, they are inclined to connect to each other [32]. In the filed of complex networks, the clustering coefficient is used to quantify the level of transitivity in a network. A high clustering coefficient means that the nodes in a real-life network tend to form triangular structures [33], which is the simplest motif can be used for sign or link prediction. Counting the number of triangular motifs is equal to the method of sign prediction based on common neighbors [34]. However, in a signed network, there is more than one single type of connection between nodes. In this study, we find that the levels of transitivity in four empirical networks are very low, which means many nodes share no common neighbors, thus we use the features of 4-node motifs to make up for the defects of 3-node motifs. Compared with the triangular structure used in the common neighbor-based method, the motif structures in Fig. 1 are more complex because the edges have both positive and negative signs. Therefore, the common neighbor-based method is a special case of SEDM, and SEDM is more general than the common neighbor-based method. As a result, SEDM has better performance in sign prediction than the common neighbor-based method.

Although all the predictors except for S_4 can improve the performance of AUC from 3.0% to 48.6% compared with the common neighbor-based method [2,7,31], there is no theoretical support and explanation for this type of intuitive methods. What is more, the other dilemma of this algorithm is that it does not distinguish the contributions of different nodes. Actually, each node in the neighborhood of the predicted edge may play a different role. Taking predictor S_1 in Fig. 3 as an example, suppose nodes M and N can form S_1 with the predicted edge (A, B). Node M can form three S_1 (i.e., the triads of (A, M, C), (C, M, D), and (B, M, D)) with three negative edges (i.e., (A, C), (C, D) and (D, B)) in Fig. 3. Different from node M, node N can form three such predictors (i.e., triads (A, N, E), (E, N, F), and (B, N, F)) with three positive edges (i.e., (A, E), (E, F), and (F, B)). This figure shows that as a neighbor node of the node pair A and B, node M promotes the formation of a negative

edge between this node pair, while node N contributes to a positive edge between this node pair. Therefore, M and N have different roles in the task of inferring the sign of predicted edge (A, B).

In summary, SEDM has two disadvantages. One is that this method is based on the intuitive understanding of local structures in a real-world network, and a theoretical explanation for the method has not yet been proposed. As a result, it is difficult to assess the reliability of it. The other is that SEDM does not distinguish the role of each node for predicting the sign of a predicted edge. In next section, we will propose a Single Motif Naive Bayes model (SMNB) to solve the above problems.

4.2. Single Motif Naive Bayes model

Although the existing naive Bayes models [16,17] can distinguish the roles of different nodes for link prediction, they have not been applied to sign prediction and are based only on 3-node motifs, in which 4-node motifs are not considered. As shown in Fig. 4(a), the role function of the 3-node motif considers the influence of node C. For 4-node motifs in Fig. 4(b) and (c), in addition to the predicted edge (A, B), there are two nodes C and D. The role function can be computed in three possible ways: (i) only consider the role of node C; (ii) only consider the role of node D; (iii) take nodes C and D as a whole, and then consider the role of the edge (C, D). The first two methods only consider the partial structures of a motif except the predicted edge, which are not comprehensive enough. Therefore, we use the third method to calculate the role function based on 4-node motifs. That is, consider the structures in the red dotted boxes as a whole in Fig. 4(b) and (c), and then calculate their roles in the prediction task.

For a given signed network $G(V, E, L)$ and a train set partition E^T, suppose (A, B) is an edge between nodes A and B, we use p_{AB} to indicate that the edge is positive, and n_{AB} to indicate that the edge is negative. The posterior probability that (A, B) is positive or negative can be calculated as:

$$P(p_{AB}|S_i(A,B)) = \frac{P(p_{AB}) \cdot P(S_i(A,B)|p_{AB})}{P(S_i(A,B))}, \tag{4}$$

$$P(n_{AB}|S_i(A,B)) = \frac{P(n_{AB}) \cdot P(S_i(A,B)|n_{AB})}{P(S_i(A,B))}, \tag{5}$$

where $S_i(A, B)$ represents a set of nodes or edges that form predictor S_i with A and B. At this time, for a given predicted edge (A, B), according to the Maximum A Posteriori Estimation [35], the sign of this edge can be determined by calculating the ratio of these two posterior probabilities to assign a score to the edge and comparing this score with 1. $\frac{P(p_{AB}|S_i(A,B))}{P(n_{AB}|S_i(A,B))} > 1$ means that $S_i(A, B)$ is caused by p_{AB}, that is, (A, B) is a positive edge. On the contrary, $\frac{P(p_{AB}|S_i(A,B))}{P(n_{AB}|S_i(A,B))} < 1$ means that $S_i(A, B)$ is caused by n_{AB}, that is, (A, B) is a negative edge. The score of edge (A, B) can be calculated as:

$$r_{AB} = \frac{P(p_{AB})}{P(n_{AB})} \cdot \frac{P(S_i(A,B)|p_{AB})}{P(S_i(A,B)|n_{AB})} = \frac{P(p_{AB})}{P(n_{AB})} \cdot \prod_{M \in S_i(A,B)} \frac{P(M|p_{AB})}{P(M|n_{AB})} = \frac{P(p_{AB})}{P(n_{AB})} \cdot \prod_{M \in S_i(A,B)} \frac{P(n_{AB})}{P(p_{AB})} \cdot \frac{P(p_{AB}|M)}{P(n_{AB}|M)}. \tag{6}$$

Here, $P(p_{AB})$ represents the prior probability that nodes A and B are linked by a positive edge, and $P(n_{AB})$ represents the prior probability that nodes A and B are linked by a negative edge. $P(p_{AB})$ and $P(n_{AB})$ can be calculated as:

$$P(p_{AB}) = \frac{X}{|E|}, \tag{7}$$

$$P(n_{AB}) = \frac{Y}{|E|}, \tag{8}$$

where $|E|$ indicates the number of all edges in the signed network, X and Y indicate the number of positive and negative edges respectively. $P(p_{AB}|M)$ denotes the probability that there is a positive edge between the pair of nodes that form the predictor S_i with node M. It can be expressed as:

$$P(p_{AB}|M) = \frac{N_{\triangle^+ S_{i_M}}}{N_{\triangle^+ S_{i_M}} + N_{\triangle^- S_{i_M}}}, \tag{9}$$

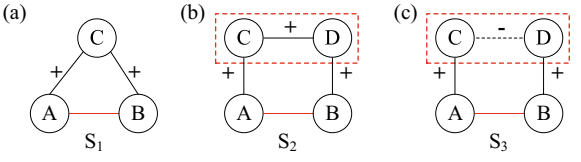

Fig. 4. The calculations of roles in different kind of motifs. (a) The role of node C in predictor S_1; (b) The role of the positive edge (C, D) in predictor S_2; (c) The role of the negative edge (C, D) in predictor S_3.

where $N_{\triangle^+S_{i_M}}$ and $N_{\triangle^-S_{i_M}}$ represent the number of positive edges and the number of negative edges in the neighborhood that form the predictor S_i with node M, respectively. Therefore, it is possible to obtain a probability that there is a negative edge between the pair of nodes that form the predictor S_i with node M:

$$P(n_{AB}|M) = 1 - P(p_{AB}|M) = \frac{N_{\triangle^-S_{i_M}}}{N_{\triangle^+S_{i_M}} + N_{\triangle^-S_{i_M}}}. \tag{10}$$

Combining Eqs. (7)–(10), r_{AB} can be simplified as:

$$r_{AB} = a^{-1} \prod_{M \in S_i(A,B)} a \frac{N_{\triangle^+S_{i_M}} + 1}{N_{\triangle^-S_{i_M}} + 1}, \tag{11}$$

where $a = \frac{P(n_{AB})}{P(p_{AB})} = \frac{Y}{X}$. To prevent the expression from being meaningless because the denominator is 0, the numerator and denominator in Eq. (10) are added 1. At this point, given a node M, its role function based on the predictor S_i can be defined as:

$$R_{S_{i_M}} = \frac{N_{\triangle^+S_{i_M}} + 1}{N_{\triangle^-S_{i_M}} + 1}. \tag{12}$$

Thus, Eq. (11) can be described as:

$$r_{AB} = a^{-1} \prod_{M \in S_i(A,B)} a R_{S_{i_M}}. \tag{13}$$

Obviously, for a given network, a is a constant. Take the logarithm of Eq. (13):

$$r'_{AB} = log\left(\prod_{M \in S_i(A,B)} a R_{S_{i_M}}\right) = \sum_{M \in S_i(A,B)} log\left(a R_{S_{i_M}}\right) = \underbrace{|S_i(A,B)|loga}_{SEDM} + \underbrace{\sum_{M \in S_i(A,B)} logR_{S_{i_M}}}_{\text{the role function of predictor } S_i}. \tag{14}$$

There are two parts when calculating the score of a predicted edge using the Single Motif Naive Bayes (SMNB) model. The first part SEDM can be calculated by the number of predictors S_i consisting of the predicted edge and its neighbor nodes (i.e., the number of edge-dependent motifs). The second part is the sum of the role of each neighbor node, which forms predictor

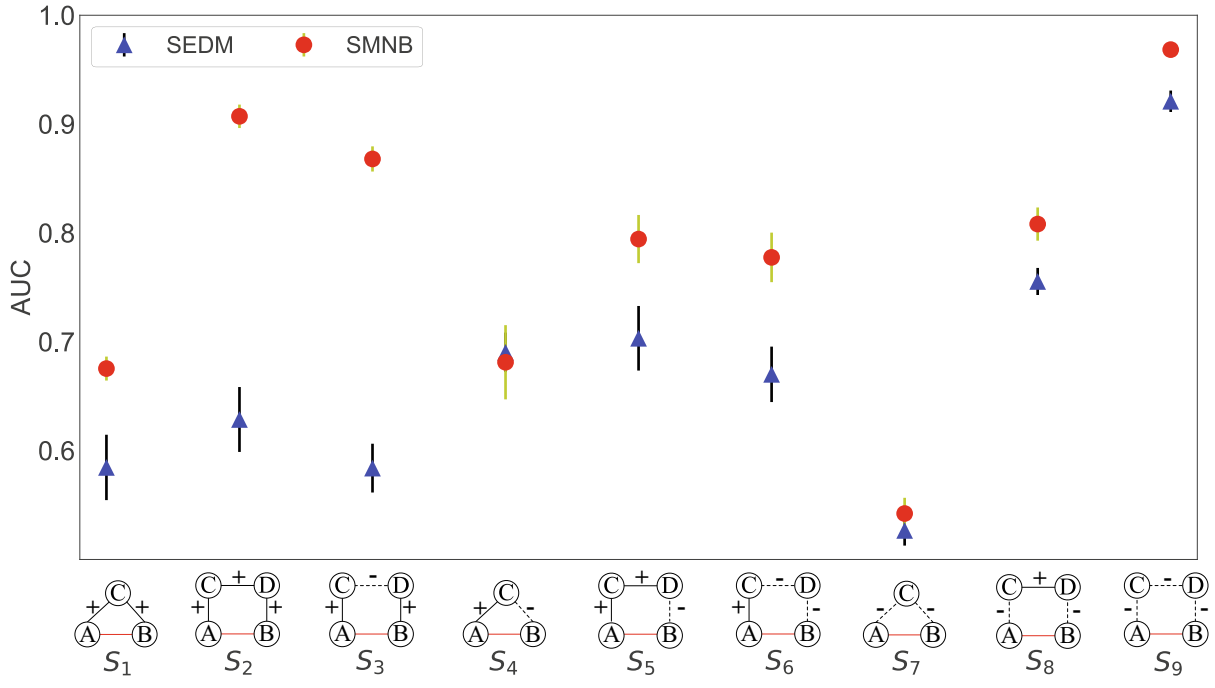

Fig. 5. The AUC values of SEDM and SMNB in the Bitcoinalpha network. SEDM represents the Single Edge-Dependent Motif based method, and SMNB represents the Single Motif Naive Bayes model.

S_i with the predicted edge. If $R_{S_{i_M}} = 1$ for each neighbor, SMNB will degenerate into SEDM in Section 4.1. That is to say, in this case we only consider the number of motifs formed by the predicted edge and other nodes in its neighborhood, regardless of the roles of these nodes.

For a given signed network G, which contains X positive edges and Y negative edges, the score of the predicted edge (A, B) can be calculated by the Single Motif Naive Bayes model. Its calculation is shown in Algorithm 2.

Algorithm 2. The Single Motif Naive Bayes Algorithm

Input: $G(V, E, S)$: the signed network
(A, B) : the predicted edge
S_i : the predictor for sign prediction
k : the order of S_i
a : the value of $\frac{Y}{X}$
Output: r'_{AB} : the score of the predicted edge (A, B)
If k equals to 3
Initialize $S_1num = S_4num = S_7num = 0$
Initialize $S_1R = S_4R = S_7R = 0$
For each node M in G
If (M, A) and (M, B) are positive edges (i.e., node M can form S_1 with (A, B)) **do**
S_1num += 1
Initialize $N_{\triangle^+S_{1_M}} = 0$, $N_{\triangle^-S_{1_M}} = 0$
For each edge (e_3, e_4) in G
If (M, e_3), (M, e_4) and (e_3, e_4) are positive edges **do**
$N_{\triangle^+S_{1_M}}$ += 1
Else If (M, e_3) and (M, e_4) are positive edges, and (e_3, e_4) is a negative edge **do**
$N_{\triangle^-S_{1_M}}$ += 1
endif
endfor
$R_{S_{1_M}} = \frac{N_{\triangle^+S_{1_M}}+1}{N_{\triangle^-S_{1_M}}+1}$
S_1R += $log\left(R_{S_{1_M}}\right)$
⋮
Else If (M, A) and (M, B) are negative edges (i.e., node M can form S_7 with (A, B)) **do**
S_7num += 1
Initialize $N_{\triangle^+S_{7_M}} = 0$, $N_{\triangle^-S_{7_M}} = 0$
For each edge (e_3, e_4) in G
If (M, e_3), (M, e_4) are negative edges and (e_3, e_4) is a positive edge **do**
$N_{\triangle^+S_{7_M}}$ += 1
Else If (M, e_3), (M, e_4) and (e_3, e_4) are all negative edges **do**
$N_{\triangle^-S_{7_M}}$ += 1
endif
endfor
$R_{S_{7_M}} = \frac{N_{\triangle^+S_{7_M}}+1}{N_{\triangle^-S_{7_M}}+1}$
S_7R += $log\left(R_{S_{7_M}}\right)$
endif
endfor
endif
If k equals to 4
Initialize $S_2num = S_3num = S_5num = S_6num = S_8num = S_9num = 0$
Initialize $S_2R = S_3R = S_5R = S_6R = S_8R = S_9R = 0$
For each edge (e_1, e_2) in G
If (e_1, e_2), (e_1, A), (e_2, B) are all positive edges or (e_1, e_2), (e_1, B), (e_2, A) are all positive edges
(i.e., edge (e_1, e_2) can form S_2 with (A, B)) **do**
S_2num += 1

(*continued on next page*)

(*continued*)

Algorithm 2. The Single Motif Naive Bayes Algorithm

Initialize $N_{\triangle^+S_{2_{(e_1,e_2)}}} = 0$, $N_{\triangle^-S_{2_{(e_1,e_2)}}} = 0$
For each edge (e_3, e_4) in G
If ((e_1, e_3), (e_2, e_4) and (e_3, e_4) are all positive edges) or ((e_1, e_4), (e_2, e_3) and (e_3, e_4) are all positive edges) **do**
$N_{\triangle^+S_{2_{(e_1,e_2)}}}$ += 1
Else If ((e_1, e_3), (e_2, e_4) are all positive edges and (e_3, e_4) is a negative edge) or ((e_1, e_4), (e_2, e_3) are all positive edges and (e_3, e_4) is a negative edge) **do**
$N_{\triangle^-S_{2_{(e_1,e_2)}}}$ += 1
endif
endfor
$R_{S_{2_{(e_1,e_2)}}} = \frac{N_{\triangle^+S_{2_{(e_1,e_2)}}}+1}{N_{\triangle^-S_{2_{(e_1,e_2)}}}+1}$
S_2R += $log\left(R_{S_{2_{(e_1,e_2)}}}\right)$
⋮
Else If (e_1, e_2), (e_1, A), (e_2, B) are all negative edges or (e_1, e_2), (e_1, B), (e_2, A) are all negative edges (i.e., edge (e_1, e_2) can form S_9 with (A, B)) **do**
S_9num += 1
Initialize $N_{\triangle^+S_{9_{(e_1,e_2)}}} = 0$, $N_{\triangle^-S_{9_{(e_1,e_2)}}} = 0$
For each edge (e_3, e_4) in G
If ((e_1, e_3), (e_2, e_4) are negative edges and (e_3, e_4) is a positive edge) or ((e_1, e_4), (e_2, e_3) are negative edges and (e_3, e_4) is a positive edge) **do**
$N_{\triangle^+S_{9_{(e_1,e_2)}}}$ += 1
Else If ((e_1, e_3), (e_3, e_4) and (e_2, e_4) are all negative edges) or ((e_1, e_4), (e_3, e_4) and (e_2, e_3) are all negative edges) **do**
$N_{\triangle^-S_{9_{(e_1,e_2)}}}$ += 1
endif
endfor
$R_{S_{9_{(e_1,e_2)}}} = \frac{N_{\triangle^+S_{9_{(e_1,e_2)}}}+1}{N_{\triangle^-S_{9_{(e_1,e_2)}}}+1}$
S_9R += $log\left(R_{S_{9_{(e_1,e_2)}}}\right)$
endif
endfor
endif
If S_i equals to S_1
$r'_{AB} = S_1num * loga + S_1R$
⋮
Else If S_i equals to S_9
$r'_{AB} = S_9num * loga + S_9R$
endif

4.3. Performance evaluation

To better demonstrate the performance of SMNB, we compare it with SEDM in four real-world networks. The result of the Bitcoinalpha network is shown in Fig. 5. The blue scatter points represent the results of using SEDM (the first part of SMNB, i.e., the number of the edge-dependent motifs) for sign prediction, and the red scatter points are the results of using the entire SMNB for sign prediction. It is shown that SMNB has a greatly higher predictive ability than SEDM, except the predictor S_4. The same conclusion can be obtained in the other three empirical networks, which indicate that the role function can improve the performance of sign prediction. Therefore, the roles of different nodes need to be distinguished.

As shown in Fig. 5, after adding the role function, 4-node predictors (i.e., S_2, S_3, S_5, S_6, S_8 and S_9) have higher performance improvements than 3-node predictors (i.e., S_1, S_4 and S_7). This indicates that the role function has greater impacts on 4-node predictors than 3-node predictors, especially predictors S_2, S_3, S_5 and S_6, and this conclusion is also applicable to the other three networks. Moreover, predictor S_9 can obtain the best performance, followed by predictor S_2. These two predictors have similar structures, that is, except the predicted edge, the other three edges in the same predictor have the same sign. Although the roles of different motifs in real-life networks are not universal, we can compare the predictive performance of these motifs to mine important motifs in different networks. Furthermore, most of the existing statistics depict the micro-scale characteristics of network topology (i.e., the properties of one node, two nodes or three nodes), but we use 4-node motifs to describe the mesoscale network properties. Therefore, there is no direct relationship between the micro-scale statistics we introduced in Section 2 and the motifs we used in this study, which is the advantage of our proposed algorithm induced from high-order statistics of real-life networks.

4.4. Time complexity analysis

Let n be the number of nodes in a real-life network and m be the number of edges. For a given predicted edge (A, B), when using the *SEDM* algorithms based on 3-node and 4-node motifs to compute the score of the edge, all the nodes and edges in the network need to be traversed, respectively. For the 5-node motifs, except the predicted edge, the algorithms search for triads consisting of two edges, which means the edges need to be traversed twice by a two-level iteration. Therefore, the time complexities of *SEDM* based on 3-node, 4-node and 5-node motifs are $O(n)$, $O(m)$ and $O(m^2)$, respectively. The *SMNB* algorithm adds a role function on the basis of *SEDM*, which means that it needs to add a layer iteration to traverse all the edges. Therefore, the time complexities of the *SMNB* algorithms based on 3-node, 4-node and 5-node motifs are $O(nm)$, $O(m^2)$ and $O(m^3)$, respectively.

5. Sign prediction by Two Motif Naive Bayes model

5.1. Two edge-dependent Motif based method

The aforementioned motif naive Bayes model is based on only a single type of motif, but the predicted edge may involve in multiple types of motifs in a real-life network. For example, when inferring the sign of the predicted edge (A, B) in Fig. 6(a), which is represented by a solid red line, there are two S_1, one S_4, two S_5 and one S_8 on the predicted edge, these different kinds of motifs can be combined for sign prediction. However, combining all types of predictors for prediction has a higher computational complexity. Therefore, in this section, we combine two different predictors for sign prediction, and several combinations of two predictors are shown in Fig. 6(b). Next, we will introduce the Two Edge-dependent Motif (TEDM) based method and the Two Motif Naive Bayes (TMNB) model, respectively.

The method of calculating the number of Two Edge-Dependent Motif (TEDM) is derived from SEDM, it can be expressed as: the score of the predicted edge is the sum of the number of these two predictors formed by the predicted edge and its neighbor nodes. That is, for an edge (A, B) whose sign needs to be predicted, if the predictors S_i and S_j are used for sign prediction, the score for this edge can be calculated as:

$$r_{AB} = |S_i(A, B)| + |S_j(A, B)|. \tag{15}$$

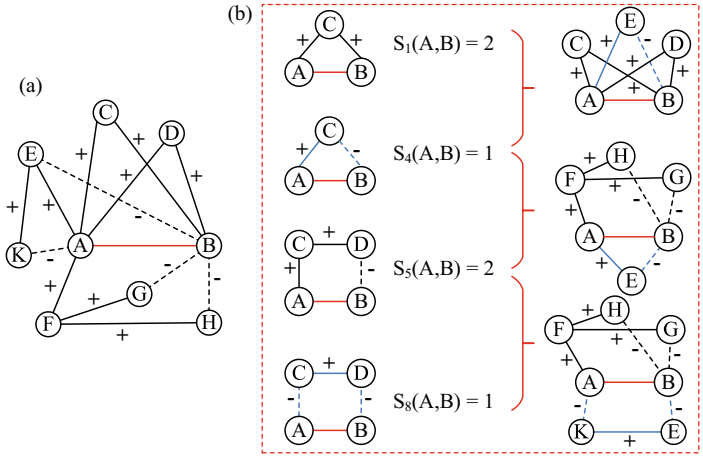

Fig. 6. The schematic diagram of sign prediction by fusing the information of two motifs.

For example, nodes C and D can form two S_1 with the predicted edge (A, B) and node E can form one S_4 with the predicted edge (A, B) in Fig. 6(a). Therefore, when using S_1 and S_4 to predict the sign of (A, B), the score of this edge is calculated as $r_{AB} = |S_1(A, B)| + |S_4(A, B)| = 2 + 1 = 3$, as shown in Fig. 6(b).

5.2. Two Motif Naive Bayes model

The result of sign prediction by a single motif shows that calculating the roles of different nodes that form a certain motif with the predicted edge can improve the predictive performance. When using two types of motifs for sign prediction, the roles of nodes that form the two kinds of motifs with the predicted edge all need to be calculated. For example, nodes C and D form two predictors S_1 with the predicted edge (A, B) in Fig. 6(a), while node E forms a predictor S_4 with the predicted edge. If predictors S_1 and S_4 are combined for sign prediction, the roles of nodes C, D and E all need to be quantified.

Suppose two predictors S_i and S_j are used for sign prediction, the posterior probability that edge (A, B) is positive or negative can be calculated as:

$$P\left(p_{AB}|\left(S_i(A,B),S_j(A,B)\right)\right)=\frac{P(p_{AB})\cdot P\left(\left(S_i(A,B),S_j(A,B)\right)|p_{AB}\right)}{P\left(S_i(A,B),S_j(A,B)\right)}=\frac{P(p_{AB})\cdot P(S_i(A,B)|p_{AB})\cdot P\left(S_j(A,B)|p_{AB}\right)}{P\left(S_i(A,B),S_j(A,B)\right)}, \tag{16}$$

$$P\left(n_{AB}|\left(S_i(A,B),S_j(A,B)\right)\right)=\frac{P(n_{AB})\cdot P\left(\left(S_i(A,B),S_j(A,B)\right)|n_{AB}\right)}{P\left(S_i(A,B),S_j(A,B)\right)}=\frac{P(n_{AB})\cdot P(S_i(A,B)|n_{AB})\cdot P\left(S_j(A,B)|n_{AB}\right)}{P\left(S_i(A,B),S_j(A,B)\right)}, \tag{17}$$

where $S_i(A, B)$ and $S_j(A, B)$ represent two sets of nodes that form predictors S_i and S_j with edge (A, B), respectively. At this time, the score of this edge can be calculated as:

$$r_{AB}=\frac{P\left(p_{AB}|\left(S_i(A,B),S_j(A,B)\right)\right)}{P\left(n_{AB}|\left(S_i(A,B),S_j(A,B)\right)\right)}=\frac{P(p_{AB})}{P(n_{AB})}\cdot\frac{P(S_i(A,B)|p_{AB})}{P(S_i(A,B)|n_{AB})}\cdot\frac{P\left(S_j(A,B)|p_{AB}\right)}{P\left(S_j(A,B)|n_{AB}\right)}, \tag{18}$$

where $\frac{P(S_i(A,B)|p_{AB})}{P(S_i(A,B)|n_{AB})}$ can be organized into the following form:

$$\frac{P(S_i(A,B)|p_{AB})}{P(S_i(A,B)|n_{AB})}=\prod_{M\in S_i(A,B)}\frac{P(M|p_{AB})}{P(M|n_{AB})}=\prod_{M\in S_i(A,B)}\frac{P(n_{AB})}{P(p_{AB})}\cdot\frac{P(p_{AB}|M)}{P(n_{AB}|M)}. \tag{19}$$

Similarly, $\frac{P\left(S_j(A,B)|p_{AB}\right)}{P\left(S_j(A,B)|n_{AB}\right)}$ can be calculated as:

$$\frac{P\left(S_j(A,B)|p_{AB}\right)}{P\left(S_j(A,B)|n_{AB}\right)}=\prod_{H\in S_j(A,B)}\frac{P(n_{AB})}{P(p_{AB})}\cdot\frac{P(p_{AB}|H)}{P(n_{AB}|H)}. \tag{20}$$

Combining Eqs. (18)–(20), r_{AB} can be simplified as:

$$r_{AB}=\frac{P(p_{AB})}{P(n_{AB})}\cdot\prod_{M\in S_i(A,B)}\frac{P(n_{AB})}{P(p_{AB})}\cdot\frac{P(p_{AB}|M)}{P(n_{AB}|M)}\cdot\prod_{H\in S_j(A,B)}\frac{P(n_{AB})}{P(p_{AB})}\cdot\frac{P(p_{AB}|H)}{P(n_{AB}|H)}. \tag{21}$$

According to the process from Eqs. (7)–(14), Eq. (21) can be simplified as:

$$r_{AB}=a^{-1}\prod_{M\in S_i(A,B)}aR_{S_{i_M}}\prod_{H\in S_j(A,B)}aR_{S_{j_H}}. \tag{22}$$

Here, $a=\frac{P(n_{AB})}{P(p_{AB})}=\frac{Y}{X}$, which is introduced in Section 4.2. $R_{S_{i_M}}=\frac{N_{\triangle+S_{i_M}}+1}{N_{\triangle-S_{i_M}}+1}$ is the role function of node M based on predictor S_i, and $R_{S_{j_H}}=\frac{N_{\triangle+S_{j_H}}+1}{N_{\triangle-S_{j_H}}+1}$ is the role function of node H based on predictor S_j. We use a logarithmic function on both sides of Eq. 22, it can be written as

$$r'_{AB}=\left(\underbrace{|S_i(A,B)|+|S_j(A,B)|}_{TEDM}\right)loga+\underbrace{\sum_{M\in S_i(A,B)}logR_{S_{i_M}}}_{\text{the role function of predictor } S_i}+\underbrace{\sum_{H\in S_j(A,B)}logR_{S_{j_H}}}_{\text{the role function of predictor } S_j}, \tag{23}$$

where $loga$ is a constant. The first part is TEDM in Section 5.1, which is based on calculating the number of two predictors formed by the predicted edge and other nodes. The second and third parts are the role functions based on predictors S_i and S_j, respectively.

Table 2
The comparison of four sign prediction methods. SEDM represents the Single Edge-Dependent Motif based method, SMNB represents the Single Motif Naive Bayes model, TEDM represents the Two Edge-Dependent Motif based method, and TMNB represents the Two Motif Naive Bayes model.

Method	Motif	Bitcoinalpha	Bitcoinotc	Wiki-RfA	Slashdot
SEDM	S_1	0.589	0.693	0.650	0.552
	S_4	0.687	0.601	0.557	0.526
	S_5	0.710	0.668	0.562	0.610
	S_8	0.787	0.717	0.723	0.757
SMNB	S_1	0.674	0.701	0.659	0.557
	S_4	0.671	0.589	0.539	0.520
	S_5	0.800	0.712	0.809	0.838
	S_8	0.798	0.723	0.855	0.836
TEDM	$S_1 + S_4$	0.555	0.630	0.618	0.534
	$S_1 + S_8$	0.754	0.739	0.723	0.763
	$S_5 + S_8$	0.778	0.684	0.586	0.684
TMNB	$S_1 + S_4$	0.762	0.732	0.686	0.574
	$S_1 + S_8$	0.828	**0.804**	0.863	0.842
	$S_5 + S_8$	**0.865**	0.774	**0.856**	**0.874**

5.3. Performance evaluation

We use TMNB to perform the sign prediction for experimental social networks, and the results are shown in Table 2. It can be found that SMNB has higher predictive performance than SEDM and TMNB outperforms TEDM, which indicates that taking into consideration role functions can improve the performance of sign prediction. Besides, although node role functions are considered in both TMNB and SMNB, TMNB has better predictive performance than SMNB because TMNB also incorporates two motifs. However, TEDM only integrates two motifs without considering role functions, which results in that TEDM is not always superior to SMNB.

We compare TMNB with the state-of-the-art approaches, including the Local Naive Bayes model (LNB), the Parallel Link Sign Prediction method (PLSP) [9] and the Deep Network Embedding with Structural Balance Preservation method (DNE-SBP) [36], for the Wiki-RfA and Slashdot networks. The LNB model can quantify the different roles of common neighbors of a node pair. The PLSP method can extract both global and local structural features. The global features are derived from status theory [22], feedback transmission theory [37] and several prior estimates from existing links; and the local features are derived from neighborhood information and balance theory [38]. Then, two speedup strategies (i.e., dataset division and feature selection) are presented to shorten the training time. The DNE-SBP method is designed to learn the low-dimensional node vector representations with structural balance preservation using a deep network embedding model in signed networks. The comparison of our proposed method and the state-of-the-art methods is shown in Fig. 7, which shows that TMNB has higher performance than LNB, PLSP and DNE-SBP methods.

Although TMNB can obtain higher prediction results than the state-of-the-art methods, it has the following shortcomings. On one hand, it uses only two types of motifs and is thus not comprehensive enough. On the other hand, any two kinds of motifs can be combined for sign prediction, randomly selecting the motifs will result in unstable prediction results. Therefore, we use all the 3-node and 4-node motifs as features for sign prediction. Specifically, the score of a predicted edge cal-

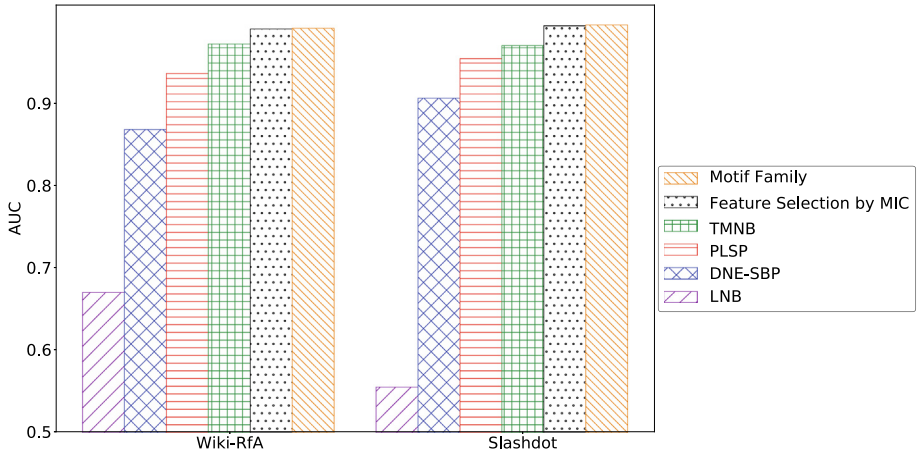

Fig. 7. The comparison of our proposed method and the state-of-the-art methods.

culated by the Single Motif Naive Bayes model (SMNB) based on a type of motif is a one-dimensional feature, and the edge scores calculated by 9 types of motifs are 9-dimensional features. And then we use the XGBoost classifier to predict the edge sign. In [39], all the networks with the same triad significance profile are classified into a class which is collectively referred to as "network superfamily". Inspired by it, we use all the types of 3-node and 4-node motifs for sign prediction, so we call them "motif family". The results are shown in Fig. 7, which indicate that compared to TMNB, PLSP, DNE-SBP and LNB, the motif family can greatly improve the performance of sign prediction.

5.4. Motif correlation detection

In the previous sections, we used different motifs for sign prediction without exploring the relationship between them. Mining the correlation among different motifs not only helps to uncover the evolution mechanism of the network, that is, those motifs that are highly correlated tend to transform each other in dynamic social networks [40,41], but also contributes to effective feature selection [42,43]. Therefore, we attempt to explore the relationship between different motifs in this section.

Generally, the Pearson correlation coefficient is a classical statistic to characterize the correlation between two variables [44,45]. It is a linear correlation coefficient that reflects the linear correlation degree of two variables, and the range of its value is between −1 and 1. The value is 1 means that the two variables are completely positively correlated, 0 means linearly independent, −1 means completely negatively correlated. The larger the absolute value, the stronger the linear relationship. The Pearson correlation coefficient metric can be calculated as

$$\rho = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right), \tag{24}$$

where $\overline{X}$ and $\overline{Y}$ are the average values of the sample values of the variables X and Y. σ_X and σ_Y are the standard deviations of the sample values of the variables X and Y, respectively. The main disadvantage of the Pearson correlation coefficient is that it can not measure the nonlinear relationship between variables, and it is susceptible to noise.

Compared with the Pearson correlation coefficient, the Maximal Information Coefficient (i.e., MIC) [18] has the following three advantages. First, it can not only get the functional relationship between variables, including linear function relation and nonlinear function relation, but also mine the non-function relation. Besides, the MIC metric can be used to compare not only the strength of the same correlation vertically, but also the strength of different relationships horizontally. Finally, the MIC metric depends only on the ordering of the data, and it is constant in the order-preserving transformation axis. The calculation process of the MIC metric can be described as follows. First, draw a grid on the scatterplot of two variables, and find the largest mutual information value. Then, normalize the largest mutual information value. Finally, select the maximum value of mutual information at different scales as the MIC value. The formula of the MIC metric can be written as

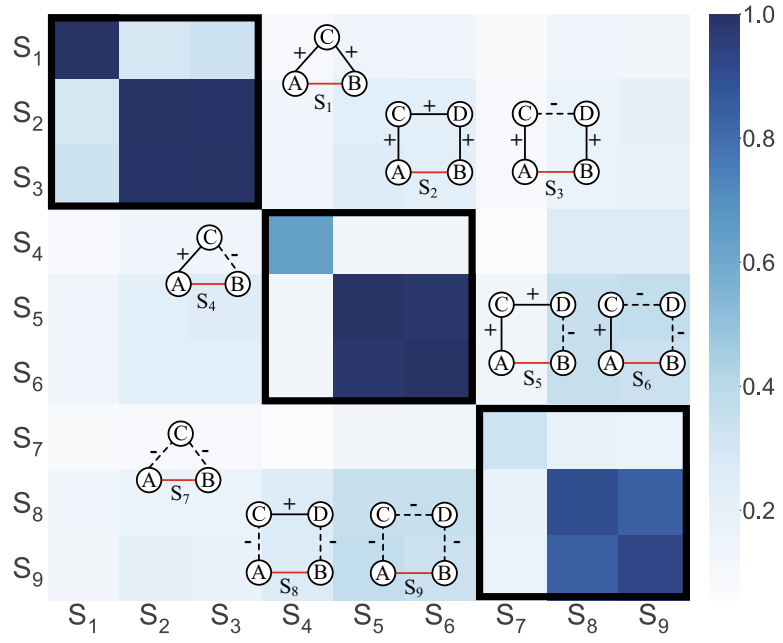

Fig. 8. The correlation matrix of Maximal Information Coefficients (MIC) of all the motif predictors.

$$MIC[X;Y] = \max_{|X||Y|<B} \frac{I[X;Y]}{\log_2(min(|X|,|Y|))}, \tag{25}$$

where $I[X;Y]$ represents the mutual information between X and Y, which is calculated as $\sum_{X,Y} p(X,Y)\log_2 \frac{p(X,Y)}{p(X)p(Y)}$. $p(X,Y)$ is the joint probability density distribution, $p(X)$ and $p(Y)$ are the marginal probability distribution functions of X and Y. $|X|$ and $|Y|$ represent the number of segments in the X and Y directions, respectively. $|X||Y| < B$ means that the total number of all squares cannot be greater than B. Generally, the value of B is 0.6 power of the total number of samples, which is an empirical value.

We use the MIC metric to measure the correlations between different motifs. The time complexity of MIC is $O(m^2)$ [46], where m represents the number of edges. The redundancy (i.e., correlation) between any two motifs S_i and S_j is defined as $MIC(S_i, S_j)$. The larger the value of $MIC(S_i, S_j)$, the stronger the substitutability between the motifs S_i and S_j (i.e., the stronger the redundancy). $MIC(S_i, S_j) = 0$ indicates that S_i and S_j are independent of each other. Then we use a heat map to color the correlations among all the predictors, where the color intensity indicates the strength of the correlation. That is to say, the darker the color, the stronger the correlation between the features. The heat map of the Bitcoinalpha network is shown in Fig. 8. In each black box, the neighbor node and the neighbor edges of the predicted edge (A, B) in these predictors link with the predicted edge by the same pattern. Therefore, three predictors in each black box can use the same type of SMNB model for sign prediction. According to the result of feature correlation in Fig. 8, there is a strong correlation between two 4-node predictors in each black box, while the correlations between 3-node and 4-node predictors are all weak. For each pair of strongly correlated 4-node motifs, the neighbor edges of the predicted edge in these two motifs are different from each other. This indicates that it is feasible to consider the two neighbor nodes of the predicted edge in the 4-node motifs as a whole, and then calculate the role of this edge. The same conclusion can be obtained in the other three real-life networks, which shows that the results of motif correlation mining are universal.

According to the correlation matrix, feature selection can be performed by removing the redundant features. Then, using the selected features for sign prediction can reduce the computation. In the correlation matrix, S_2 and S_3, S_5 and S_6, S_8 and S_9 are strongly correlated, while three 3-node motifs (S_1, S_4 and S_7) are independent of each other. Since the redundancy is great when the motifs are strongly correlated, we choose S_1, S_2, S_4, S_5, S_7 and S_8 for sign prediction to reduce the redundancy, and the results of the Wiki-RfA and Slashdot networks are shown in Fig. 7. The result of feature selection by MIC correlation matrix is very close to that of the motif family, which indicates the correlation matrix can effectively filter the redundant features so as to reduce the computation on the premise that the performance drops little.

6. Conclusion

In summary, we proposed a Single Motif Naive Bayes (SMNB) model to predict the edge signs in online social networks. This model not only explains the prediction mechanism of the Single Edge-Dependent Motif based method (SEDM), but also considers the roles of different nodes (for 3-node predictors) and edges (for 4-node predictors) when using multiple motif information for sign prediction. The sign prediction performance on four empirical signed networks indicates that SMNB has a greater predictive ability than SEDM (i.e., calculating the motif number on the predicted edge). Next, we extended the SMNB model, and proposed a Two Motif Naive Bayes (TMNB) model. Experimental results on real-world networks validated the superiority of the proposed method comparing with the state-of-the-art methods. Finally, according to the matrix of the Maximal Information Coefficient (MIC) between all the motif predictors, the intrinsic relationship between any pair of motifs is discovered.

Different kinds of motifs can be integrated into the naive Bayes model for sign prediction, which extends the traditional motif theory [47]. We proved that at the mesoscale it is reasonable to use the motifs that are distributed on a predicted edge to perform sign prediction, while the traditional motif theory only pays attention to whether the number of motifs significant in the whole network compared with its corresponding randomized version at the macro-scale [47]. Moreover, we found that the role of a node and its contribution are significantly related to the number of the special functional units it participates in, and distinguishing the roles of different nodes helps to further understand the evolution of network structures and design new algorithms to improve the performance of sign prediction. In the future, we will explore the detailed relationship between different motifs to uncover the evolution mechanism of signed social networks. Furthermore, we will combine motifs with the bayesian linear regression model [48] to perform edge weight prediction [49] in signed social networks, and select features according to the correlation between motifs [9,50].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Si-Yuan Liu: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing - original draft. **Jing Xiao:** Validation, Writing - review & editing. **Xiao-Ke Xu:** Conceptualization, Methodology, Supervision, Formal analysis, Writing - review & editing.

Acknowledgments

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

References

[1] D. Cartwright, F. Harary, Structural balance: a generalization of Heider's theory, The Psychological Review 63 (1956) 277–293.
[2] J. Leskovec, D. Huttenlocher, J. Kleinberg, Predicting positive and negative links in online social networks, in: Proceedings of the 19th International Conference on World Wide Web, USA, 2010, pp. 641–650.
[3] Q.V. Dang, C.L. Ignat, Link-sign prediction in dynamic signed directed networks, in: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), 2018, pp. 36–45.
[4] W. Yuan, K. He, D. Guan, G. Han, Edge-dual graph preserving sign prediction for signed social networks, IEEE Access 5 (99) (2017) 19383–19392.
[5] R. Guha, R. Kumar, P. Raghavan, A. Tomkins, Propagation of trust and distrust, in: Proceedings of the 13th International Conference on World Wide Web, 2004, pp. 403–412.
[6] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, S. Albayrak, Spectral analysis of signed graphs for clustering, prediction and visualization, in: SIAM International Conference on Data Mining, 2010, pp. 559–570.
[7] K.Y. Chiang, N. Natarajan, A. Tewari, I.S. Dhillon, Exploiting longer cycles for link prediction in signed networks, in: ACM International Conference on Information & Knowledge Management, 2011, pp. 1157–1162.
[8] P. Agrawal, V.K. Garg, R. Narayanam, Link label prediction in signed social networks, in: International Joint Conference on Artificial Intelligence, 2013, pp. 2591–2597.
[9] J. Zhou, L. Han, Y. Yao, X. Zeng, F. Xu, A parallel approach to link sign prediction in large-scale online social networks, Computer Journal 57 (7) (2014) 1092–1104.
[10] K. Akilal, H. Slimani, M. Omar, A very fast and robust trust inference algorithm in weighted signed social networks using controversy, eclecticism, and reciprocity, Computers & Security 83 (2019) 68–78.
[11] X. Xu, J. Zhang, S. Michael, Superfamily phenomena and motifs of networks induced from time series, Proceedings of the National Academy of Sciences of the United States of America 105 (50) (2008) 19601–19605.
[12] A. Papaoikonomou, M. Kardara, K. Tserpes, T.A. Varvarigou, Predicting edge signs in social networks using frequent subgraph discovery, IEEE Internet Computing 18 (5) (2014) 36–43.
[13] D. Song, D.A. Meyer, Link sign prediction and ranking in signed directed social networks, Social Network Analysis & Mining 5 (1) (2015) 1–14.
[14] Y.-F. Liu, T. Li, X.-K. Xu, Link prediction by multiple motifs in directed networks, IEEE Access 8 (2019) 174–183.
[15] S. Liu, J. Xiao, X.-K. Xu, Link prediction in signed social networks: from status theory to motif families, IEEE Transactions on Network Science and Engineering, doi: 10.1109/TNSE.2019.2951806 .
[16] Z. Liu, Q. Zhang, L. Lü, T. Zhou, Link prediction in complex networks: a local naive bayes model, EPL 96 (2011) 48007.
[17] J. Wu, G. Zhang, Y. Ren, X. Zhang, Q. Yang, Weighted local naive Bayes link prediction, Journal of Information Processing Systems 13 (4) (2017) 914–927.
[18] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, M.V. Gilean, P.J. Turnbaugh, E.S. Lander, M. Michael, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518.
[19] S. Kumar, F. Spezzano, V. Subrahmanian, C. Faloutsos, Edge weight prediction in weighted signed networks, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 221–230.
[20] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, V. Subrahmanian, Rev2: Fraudulent user prediction in rating platforms, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 333–341.
[21] R. West, H.S. Paskov, J. Leskovec, C. Potts, Exploiting social network structure for Person-to-Person sentiment analysis, Transactions of the Association for Computational Linguistics 2 (2014) 297–310.
[22] J. Leskovec, D. Huttenlocher, J. Kleinberg, Signed networks in social media, in: Conference on Human Factors in Computing Systems, 2010, pp. 1361–1370 .
[23] E.J.M. Newman, Assortative mixing in networks, Physical Review Letters 89 (20) (2002) 208701.
[24] R. Donner, J. Donges, Visibility graph analysis of geophysical time series: potentials and possible pitfalls, Donner, Reik; Donges, Jonathan 60 (2012) 589.
[25] K. Dong, A.R. Benson, D. Bindel, Network density of states, in: The 25th ACM SIGKDD International Conference, 2019.
[26] J. Zeng, K. Zhou, X. Ma, F. Zou, H. Wang, Exploiting cluster-based meta paths for link prediction in signed networks, in: The 25th ACM Conference on Information and Knowledge Management, 2016, pp. 1905–1908.
[27] A. Javari, H. Qiu, E. Barzegaran, M. Jalili, K. Chang, Statistical link label modeling for sign prediction: smoothing sparsity by joining local and global information, in: 17th IEEE International Conference on Data Mining, 2017, pp. 1039–1044.
[28] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, 2016.
[29] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology 143 (1) (1982) 29–36.
[30] Q. Zhang, L. Lü, W. Wang, T. Zhou, Potential theory for directed networks, Plos One 8 (2) (2013) e55437.
[31] A. Papaoikonomou, M. Kardara, K. Tserpes, T.A. Varvarigou, Predicting edge signs in social networks using frequent subgraph discovery, IEEE Internet Computing 18 (5) (2014) 36–43.
[32] M.E.J. Newman, Networks, Oxford University Press, 2018.
[33] X. Feng, J.C. Zhao, K. Xu, Link prediction in complex networks: a clustering perspective, European Physical Journal B 85 (1) (2012) 3–11.
[34] Z.-M. Ren, A. Zeng, Y.-C. Zhang, Structure-oriented prediction in complex networks, Physical Reports 750 (2018) 1–51.
[35] D. Greig, B. Porteous, A. Seheult, Exact maximum a posteriori estimation for binary images, Journal of the Royal Statistical Society 51 (1989) 271–279.
[36] X. Shen, F. Chung, Deep network embedding for graph representation learning in signed networks, IEEE Transactions on Cybernetics (2018) 1–8.
[37] C.D. Fisher, Transmission of positive and negative feedback to subordinates: a laboratory investigation, Journal of Applied Psychology 64 (5) (1979) 533–540.
[38] F. Heider, Attitudes and cognitive organization, The Journal of Psychology 21 (1946) 107–112.

[39] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, Superfamilies of evolved and designed neworks, Science (New York, N.Y.) 303 (2004) 1538–1542 .
[40] L.K. Gallos, D. Rybski, F. Liljeros, S. Havlin, H.A. Makse, How people interact in evolving online affiliation networks, Physical Review X 2 (3) (2012) 031014–031024.
[41] K. Juszczyszyn, K. Musial, M. Budka, Link prediction based on subgraph evolution in dynamic social networks, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 27–34.
[42] Z. Chen, K.Y. Chai, S.L.F. Bu, C.T. Lau, Combining mic feature selection and feature-based mspca for network traffic anomaly detection, in: Third International Conference on Digital Information Processing, 2016, pp. 176–181.
[43] G. Sun, J. Li, J. Dai, Z. Song, F. Lang, Mic-based feature selection method for iot data processing, Future Generation Computer Systems 89 (2018) 606–616.
[44] J. Gravier, V. Vignal, S. Bissey-Breton, J. Farre, The use of linear regression methods and pearson's correlation matrix to identify mechanical-physical-chemical parameters controlling the micro-electrochemical behaviour of machined copper, Corrosion Science 50 (10) (2008) 2885–2894.
[45] A.M. Gadermann, M. Guhn, B.D. Zumbo, Estimating ordinal reliability for likert-type and ordinal item response data: a conceptual, empirical, and practical guide, Practical Assessment Research & Evaluation 17 (1) (2012) 1–13.
[46] S. Wang, Y. Zhao, Y. Shu, H. Yuan, J. Geng, S. Wang, Fast search local extremum for maximal information coefficient (mic), Journal of Computational & Applied Mathematics 327 (2017) 372–387.
[47] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: Simple building blocks of complex networks, Science (New York, N. Y.) 298 (5594) (2002) 824–827 .
[48] Shiyong Cui, O. Meynberg, P. Reinartz, Bayesian linear regression for crowd density estimation in aerial images, in: 2017 Joint Urban Remote Sensing Event (JURSE), 2017, pp. 1–4 .
[49] S. Kumar, F. Spezzano, V.S. Subrahmanian, C. Faloutsos, Edge weight prediction in weighted signed networks, in: IEEE International Conference on Data Mining, 2017, pp. 221–230.
[50] J.G. Dy, C.E. Brodley, Feature selection for unsupervised learning, Journal of Machine Learning Research 5 (4) (2004) 845–889.