# SSBM: A Signed Stochastic Block Model for Multiple Structure Discovery in Large-Scale Exploratory Signed Networks

Yang Li<sup>c</sup>, Bo Yang<sup>a,b</sup>, Xuehua Zhao<sup>d</sup>, Zhejian Yang<sup>a,e</sup> and Hechang Chen<sup>a,e</sup>

<sup>a</sup>Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Changchun, 130012, China

<sup>b</sup>College of Computer Science and Technology, Jilin University, Changchun, 130012, China

<sup>c</sup>Aviation University of Air Force, Changchun, 130062, China

<sup>d</sup>School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen, 518055, China

<sup>e</sup>School of Artificial Intelligence, Jilin University, Changchun, 130012, China

# ARTICLE INFO

Keywords: Signed network Multiple structure discovery Stochastic block model Model selection

## ABSTRACT

Signed network structure discovery has received extensive attention and has become a research focus in the field of network science. However, most of the existing studies are focused on the networks with a single structure, e.g., community or bipartite, while ignoring multiple structures, e.g., the coexistence of community and bipartite structures. Furthermore, existing studies were faced with challenge regarding large-scale signed networks due to their high time complexity, especially when determining the number of clusters in the observed network without any prior knowledge. In view of this, we propose a mathematically principled method for signed network multiple structure discovery named the Signed Stochastic Block Model (SSBM). The SSBM can capture the multiple structures contained in signed networks, e.g., community, bipartite, and coexistence of them, by adopting a probabilistic model. Moreover, by integrating the minimum message length (MML) criterion and component-wise EM (CEM) algorithm, a scalable learning algorithm that has the ability of model selection is proposed to handle large-scale signed networks. By comparing state-of-the-art methods on synthetic and real-world signed networks, extensive experimental results demonstrate the effectiveness and efficiency of SSBM in discovering large-scale exploratory signed networks with multiple structures.

# 1. Introduction

Structure discovery for signed networks with positive and negative edges has received extensive attention and has become a research focus in the field of network science [1, 2, 3, 4]. Different from unsigned networks containing only one kind of edge describing a homogenous relationship [5, 6, 7], the positive edges in signed networks usually represent trust, like, or support relationships, while the negative edges usually represent distrust, dislike, or oppose relationships [8, 9]. Therefore, signed networks can characterize different relationships between individuals by adding positive and negative signs.

In recent years, researchers have found that the community, i.e., a dense subnetwork, is one of the most common structures in real-world networks [10]. Subsequently, some community discovery methods have been proposed and can be classified into two categories: discriminant and principled methods. For discriminant methods, optimization objectives, e.g., modularity, or heuristics, e.g., random walk model, should be predefined according to specific network characteristics [11, 12, 13, 14, 15, 16, 17]. However, discriminant methods are inflexible in practical applications due to the difficulty of designing appropriate objective functions. The principled methods are usually used to detect structures in signed networks because probability models can capture the intrinsic features of different structures when fitting the observed networks [3, 18]. Despite the success that previous studies achieved in structure discovery, there are two challenges remain unsolved: 1) Generalization: both discriminant methods and principled methods can mostly discover one single structure, e.g., community or bipartite, but cannot detect multiple structures, e.g., the coexistence of community and bipartite. 2) Scalability: most principled methods have high time complexity due to parameter estimation and model selection, i.e., determining the number of clusters K. To select an optimal model, these methods have to traverse all possible K values and then calculate the parameters of all the possible models [4, 18], leading to high time complexity for a slightly largescale network.

In view of this, a generalized model for characterizing multiple structures and a scalable learning algorithm for largescale exploratory signed networks are proposed in this paper. The contributions are summarized as follows:

1) A new reparameterized signed stochastic block model, namely SSBM, is proposed to characterize the multiple structures in the signed networks.

In the SSBM, a new parameter  $\Lambda$  is introduced to reparameterize the parameter  $\Pi$  in the standard SBM. The reparameterized model continues the idea of using block structure to characterize network structure, and the property of structural equivalence in the standard SBM, i.e., the nodes in the same block have similar connection patterns. This allows SSBM to characterize a single structure, e.g., community or bipartite, and even more complex multiple structures, e.g., their coexistence in the signed networks. In addition, the reparameterization can fundamentally solve the high time complexity problem encountered by most existing

<sup>\*</sup>Corresponding authors: Bo Yang and Hechang Chen

ORCID(s): 0000-0003-1927-8419 (B. Yang); 0000-0001-7835-9556 (H. Chen)

SBM learning methods, e.g.,  $O(K^2n^2)$  when K is known; otherwise,  $O(n^5)$ . This makes it possible to detect and analyze multiple structures of large-scale signed networks.

2) A scalable learning algorithm SSBM with model selection ability is proposed for large-scale exploratory signed networks.

Model selection ability indicates whether an algorithm can discover network structures without prior knowledge. For most existing signed network structure discovery methods, a serial learning mechanism is usually employed to perform parameter estimation and model selection alternately in the model space. The time complexity resulting from this mechanism is generally  $O(n^5)$ . For this purpose, a scalable learning algorithm SSBM by integrating the minimum message length (MML) [19] and component-wise EM (CEM) algorithm [20] is proposed. The SSBM can synchronously perform parameter estimation and model selection in the block space  $[K_{min}, K_{max}]$ , which can effectively reduce the learning time from  $O(n^5)$  to  $O(n^3)$ . This parallel learning mechanism is crucial for realizing multiple structure discovery of large-scale exploratory signed networks.

3) Experimental results on synthetic and real-world signed networks in terms of generalization, robustness, and scalability demonstrate superiority of SSBM.

For generalization, the experimental results on various networks validate the effectiveness of SSBM in discovering multiple structures. For robustness, according to the learned parameters  $\Lambda$ , the densities of different noises, i.e., negative edges within a block and positive edges between blocks, can be denoted explicitly. Therefore, the SSBM can overcome the influence of different noise types and densities and accurately discover the multiple structures in the signed networks. For scalability, the experimental results on large-scale synthetic and real-world signed networks confirm that SSBM can effectively handle networks with tens of thousands of nodes in minute-level time. This significant advantage allows SSBM to discover multiple structures of large-scale exploratory signed networks.

The rest of the paper is organized as follows. The proposed model and learning algorithm are described in detail in Section 2. The experimental results to validate the effectiveness and efficiency of SSBM are presented in Section 3. Section 4 summarizes the related works in terms of discriminant and principled methods, and we conclude the paper in Section 5.

# 2. Methodology

In this section, a novel signed stochastic block model, namely SSBM, is introduced by reparameterizing the standard SBM to discover multiple structures in signed networks. Then, a scalable learning algorithm is proposed to perform parameter estimation and model selection simultaneously.

## 2.1. Signed Stochastic Block Model

The standard SBM that only discovers structures contained in unsigned networks can be formalized as [21]:

$$X = (K, Z, \Phi, \Pi) \tag{1}$$

where *K* denotes the number of blocks in a network consisting of *n* nodes. The latent variable *Z* is a  $n \times K$  dimensional matrix, and  $z_{ik} = 1$  if node *i* is allocated to block *k*; otherwise,  $z_{ik} = 0$ .  $\Phi = (\phi_1, \dots, \phi_K)$  is a *K* dimensional vector, and  $\phi_k$  represents the probability of allocating a node to block *k*.  $\Pi$  is a  $K \times K$  dimensional matrix, and the element  $\pi_{kq}$  represents the probability of generating an edge between the nodes in block *k* and block *q*.

According to the above definition, the standard SBM only can characterize unsigned network structure due to matrix  $\Pi$ . For this problem, a new parameter  $\Lambda$  is introduced to reparameterize  $\Pi$ , and  $\Lambda$  represents the probability matrix that positive, negative, or null edges will be generated from a block to a node. Then a new signed stochastic block model is presented to characterize multiple structures in the signed networks at a finer granularity.

Assuming that  $E_{n \times n}$  is an adjacent matrix of an observed signed network N containing n nodes. The element  $e_{ij}$  equals to 1, -1, or 0, denoting a positive, negative, or null edge existing between node *i* and node *j*. Then, a new signed stochastic block model, named SSBM, can be formulated as follows:

$$X = (K, Z, \Phi, \Lambda) \tag{2}$$

where the parameters  $K, Z, \Phi$  are identical to them in the standard SBM, i.e., K denotes the number of blocks, Z is a  $n \times K$  dimensional latent variable indicating node assignments, and  $\Phi$  is a K dimensional vector representing probability distribution of allocating a node to different blocks.  $\Lambda$  is a  $K \times n \times 3$  dimensional block-to-node connection matrix, and the element  $\lambda_{kj} = (\lambda_{kj1}, \lambda_{kj2}, \lambda_{kj3})$  represents probability generating a positive, negative, or null edge from a node in block k and node j, respectively. Fig.1 shows a graphical model of SSBM.



Figure 1: Graphical model of SSBM.

Note that SSBM is also a generative model that can generate various signed networks with block structures. A signed network can be generated by the following steps:

Step I: Allocate a node *i* to a block *k* according to the probability  $\phi_k$ ;

Step II: If node *i* is in block *k*, then generating a positive, negative, or null edge between node *i* and node *j* according to multinomial distribution  $\lambda_{kj} = (\lambda_{kj1}, \lambda_{kj2}, \lambda_{kj3})$ .



**Figure 2:** The steps of SSBM generating signed networks. The solid lines between two nodes denote the generated positive or negative edges labeled as "+" or "-", respectively. The dotted lines denote edges to be generated, and the signs of edges are determined by multinomial distribution with the parameter  $\lambda_{kj}$  formalized as  $Mu(1, \lambda_{kj} = \{\lambda_{kj1}, \lambda_{kj2}, \lambda_{kj3}\})$ .

The graphical model of SSBM and the steps of SSBM generating signed networks as mentioned above are exhibited in Fig.1 and Fig.2, respectively. It is clear that  $e_{ij}$  is dependent on  $Z_i$  and  $\Lambda$ , and  $Z_i$  is only dependent on  $\Phi$ . Therefore, the likelihood of an observed signed network N and the latent variable Z, namely, the complete data likelihood  $p(N, Z | K, \Phi, \Lambda)$ , can be formalized as follows:

$$p(N, Z|K, \Phi, \Lambda) = p(Z|K, \Phi)p(N|K, Z, \Lambda)$$
(3)

For Equation (3),

$$p(Z|K,\Phi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \phi_{k}^{z_{ik}}$$
(4)

$$p(N|K, Z, \Lambda) = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{n} \prod_{h=1}^{3} \lambda_{kjh}^{z_{ik}\delta(a_{ij}, 2-h)}$$
(5)

where  $h \in \{1, 2, 3\}$ , and when x = y,  $\delta(x, y) = 1$ ; otherwise,  $\delta(x, y) = 0$ . Then the log form of Equation (3), i.e., the complete data log likelihood is:

$$\log p(N, Z | K, \Phi, \Lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \phi_k$$
  
+ 
$$\sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{h=1}^{3} z_{ik} \delta(a_{ij}, 2-h) \log \lambda_{kjh}$$
 (6)

# 2.2. Scalable Learning Algorithm

By combining MML criterion with CEM algorithm, a scalable learning algorithm is proposed to carry out parameter estimation and model selection simultaneously. The CEM algorithm is used to learn the parameters of blocks, and MML is used to evaluate and select blocks. If one block is of poor quality, e.g., empty, then it is discarded and no longer computed in the subsequent iterations until convergence. Different from most of the existing methods, our proposed learning algorithm focuses on block space rather than model space. The time complexity can be reduced when selecting models, and the efficiency can be improved significantly.

#### 2.2.1. Cost Function of SSBM

The detailed deduction of SSBM cost function based on MML criterion is presented in this section. Specificially, the cost function of MML is [22, 23]:

$$C(N,g) = -\log p(N|g) - \log p(g) + \frac{1}{2} \log |\mathbf{I}(g)| + \frac{d}{2}(1 + \log \kappa_d)$$
(7)

where *N* represents the observed signed network, and *g* denotes model parameters, i.e.,  $(K, \Phi, \Lambda)$  in SSBM. *d* is the dimension of *g*,  $\kappa_d \approx (2\pi e)^{-1}$  when *d* is large.  $\mathbf{I}(g) \equiv -\mathbb{E}_{p(N|g)}[D_g^2 \log p(N|g)]$  is the Fisher information matrix,  $\mathbb{E}$  and  $|\mathbf{I}(g)|$  denote expectation and determinant, respectively.

Since I(g) cannot be calculated analytically, an upper bound of I(g) is constructed by the Fisher information matrix of complete data likelihood:

$$\mathbf{I}_{c}(g) \equiv -\mathbb{E}_{p(N,Z|g)}[D_{g}^{2}\log p(N,Z|g)]$$
(8)

where  $D_g^2$  is the second derivative on g. The C(N, g) can be optimized by minimizing  $\mathbf{I}_{c}(g)$  [24].

We let  $g = (K, \Phi, \Lambda)$ ,  $\mathbf{I}_{c}(g)$ , a 3Kn + K dimensional diagonal matrix where diagonal elements are the second partial derivatives of  $\log p(N, Z|g)$ . According to Equation (8), we have:

$$-\mathbb{E}_{p(N,Z|g)}\left[\frac{\partial^2 \log p(N,Z|g)}{\partial \phi_k^2}\right] = n\phi_k^{-1} \tag{9}$$

$$-\mathbb{E}_{p(N,Z|g)}\left[\frac{\partial^2 \log p(N,Z|g)}{\partial \lambda_{kjh}^2}\right] = n\phi_k \lambda_{kjh}^{-1} \qquad (10)$$

Then the determinant of the 3Kn + K dimensional diagonal matrix  $\mathbf{I}_{c}(g)$  is as follows:

$$|\mathbf{I}_{c}(g)| = n^{3nK+K} \prod_{k=1}^{K} \phi_{k}^{-1} \prod_{q=1}^{K} \prod_{j=1}^{n} \prod_{h=1}^{3} \phi_{q} \lambda_{qjh}^{-1}$$
(11)

A noninformative prior is used to formally characterize p(g) due to  $\Phi$  independent on  $\Lambda$ .

$$p(g) = p(\phi_1, ..., \phi_k) \prod_{k=1}^{K} \prod_{j=1}^{n} p(\lambda_{kj1}, \lambda_{kj2}, \lambda_{kj3})$$
(12)

The standard Jeffrey prior [25] is adopted to characterize  $p(\phi_k, \cdots, \phi_K)$  and  $p(\lambda_{kj1}, \lambda_{kj2}, \lambda_{kj3})$ .

$$p(\phi_1, ..., \phi_k) \propto (\prod_{k=1}^K \phi_k)^{-\frac{1}{2}}$$
 (13)

$$p(\lambda_{kj1}, \lambda_{kj2}, \lambda_{kj3}) \propto (\lambda_{kj1}\lambda_{kj2}\lambda_{kj3})^{-\frac{1}{2}}$$
(14)

According to Equations (12), (13), and (14),  $-\log p(g)$  in Equation (7) can be formalized as follows:

$$-\log p(g) = -\frac{1}{2} \sum_{k=1}^{K} \log \phi_k^{-1} - \frac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{h=1}^{3} \log \lambda_{kjh}^{-1}$$
(15)

Finally, the cost function of SSBM is:

$$C(N,g) = -\log p(N|g) + \frac{K(c+1)}{2}\log n + \frac{c}{2}\sum_{k=1}^{K}\log\phi_k + \frac{d}{2}(1+\log\kappa_d)$$
(16)

-- /

where c is the number of parameters in a single block, and g denotes the model parameters, i.e.,  $(K, \Phi, \Lambda)$ .

From the perspective of information coding, Equation (16) is essentially the sum of data coding lengths, i.e.,  $-\log p(N|g)$ , Finally, the probability generating a positive, negative, or and model coding length, i.e.,  $\frac{K(c+1)}{2}\log n + \frac{c}{2}\sum_{k=1}^{K}\log \phi_k +$  null edge from block k to node j, i.e.,  $\lambda_{kj1}$ ,  $\lambda_{kj2}$ , and  $\lambda_{kj3}$ ,

 $\frac{d}{2}(1 + \log \kappa_d)$ . Optimizing the cost function is to minimize information coding length. Because the parameters of an empty block k, i.e.,  $\phi_k = 0$ , have no effect on total information coding length, the final formalization of Equation (16) can be rewritten by defining  $K_{ne} < K$  to denote the number of nonempty blocks:

$$C(N,g) = -\log p(N|g) + \frac{K_{ne}(c+1)}{2}\log n + \frac{c}{2}\sum_{\phi_k>0}\log \phi_k + \frac{d}{2}(1+\log \kappa_d)$$
(17)

#### 2.2.2. CEM-based Optimization

In this subsection, the CEM algorithm is adopted to learn model parameters in Equation (17). Obviously, the right side of Equation (17) is opposite number of the sum of log likelihood, i.e.,  $\log p(N|g)$ , and model prior, i.e.,  $-\frac{K_{ne}(c+1)}{2}\log n$  $\frac{c}{2} \sum_{\phi_k > 0} \log \phi_k - \frac{d}{2} (1 + \log \kappa_d)$ , which is equal to model posterior. The original optimization task can be transformed from minimizing the cost function to maximizing the posterior. Specifically, the learning algorithm consists of the following two steps:

**E-step:** When N and  $o^{(t-1)}$  are known, where o represents  $(\Phi, \Lambda)$  and t is the times of iterations, the expectation of complete data log likelihood, i.e., the O function:

$$Q(o, o^{(t-1)}) = \mathbb{E}_{Z}[\log p(N, Z | K, o^{(t-1)})]$$
  
=  $\sum_{i=1}^{n} \sum_{k=1}^{K} \zeta_{ik} \log \phi_{k} + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{h=1}^{3} \zeta_{ik} \delta(a_{ij}, 2-h) \log \lambda_{kjh}$   
(18)

where  $\zeta_{ik} = \mathbb{E}[z_{ik}; o^{(t-1)}]$  is the posterior probability that node *i* is allocated to block *k* when  $o^{(t-1)}$  is known, and can be calculated as follows:

$$\zeta_{ik} = \frac{\phi_k^{(t-1)} \prod_{j=1}^n \prod_{h=1}^3 \lambda_{kjh}^{\delta(a_{ij},2-h)}}{\sum_{l=1}^K \phi_l^{(t-1)} \prod_{j=1}^n \prod_{h=1}^3 \lambda_{ljh}^{\delta(a_{ij},2-h)}}$$
(19)

**M-step:** Maximizing  $Q(o, o^{(t-1)}) + \log p(o)$ , where  $\log p(o) = -\frac{K_{ne}(c+1)}{2} \log n - \frac{c}{2} \sum_{k:\phi_k>0} \log \phi_k - \frac{K_{ne}(c+1)}{2} (1 + \log \kappa_d)$ . Since  $\sum_{k=1}^{K} \phi_k = 1$ , the Laplace function to be maximized is as follows:

$$J = Q(o, o^{(t-1)}) + \log p(o) + \eta(\sum_{k=1}^{K} \phi_k - 1)$$
(20)

Then an explicit solution on  $\phi$  can be obtained by calculating the partial derivative of Equation (20).

$$\phi_k^{(t)} = \frac{\max\{0, \sum_{i=1}^n \zeta_{ik} - \frac{c}{2}\}}{\sum_{l=1}^K \max\{0, \sum_{i=1}^n \zeta_{il} - \frac{c}{2}\}}$$
(21)

can be represented as follows:

$$\lambda_{kj1} = \frac{\sum_{i=1}^{n} \zeta_{ik} \delta(a_{ij}, 1)}{\sum_{i=1}^{n} \zeta_{ik}}$$

$$\lambda_{kj2} = \frac{\sum_{i=1}^{n} \zeta_{ik} \delta(a_{ij}, -1)}{\sum_{i=1}^{n} \zeta_{ik}}$$

$$\lambda_{kj3} = \frac{\sum_{i=1}^{n} \zeta_{ik} \delta(a_{ij}, 0)}{\sum_{i=1}^{n} \zeta_{ik}}$$
(22)

In the SSBM model, the ability of model selection can be reflected by Equation (21). Specifically,  $\zeta_{ik}$  in the numerator is posterior probability allocating node *i* to block *k*, then  $\sum_{i=1}^{n} \zeta_{ik}$  can be regarded as the number of nodes allocated to block *k*. When the numerator is 0, i.e.,  $\sum_{i=1}^{n} \zeta_{ik} < \frac{c}{2}$ , block *k* will be discarded due to be empty, and the parameters are not estimated in subsequent iterations. For SSBM, all blocks are empty, and the algorithm is invalid when *c* is equal to 3*n*. However, concerning on the right side of Equation (17), it contains the log form of model posterior, which consists of a log form of the Dirichlet prior of  $\phi_k$ , i.e.,  $\frac{c}{2} \sum_{\phi_k>0} \log \phi_k$ , and a log likelihood, i.e.,  $\log p(N|g)$ , when neglecting constant term.

$$p(\phi_1, \dots, \phi_K) \propto exp\{-\frac{c}{2}\sum_{k=1}^K log\phi_k\} = \prod_{k=1}^K \phi_k^{-\frac{c}{2}}$$
 (23)

In Equation (23),  $-\frac{c}{2}$  is a parameter of the Dirichlet prior, and it has little influence on detecting results of network structure when the data of observable signed network is sufficient. That is, varying *c* in Equation (21) will not have a significant impact on the posterior of  $\Phi$ . Moreover, the standard SBM needs *K* parameters to characterize a block, while SSBM requires 3*n* parameters. The SSBM is essentially an extension of the standard SBM from unsigned networks to signed networks, and will degenerate to the standard SBM regardless of the signs of edges, and then the number of parameters of each block in SSBM is also *K*. Therefore,  $c/2 = K_{ne}$ , i.e., the number of nonempty blocks, can be set in the real applications.

Note that an important heuristic information on setting the upper bound of detectable block space, i.e.,  $K_{max}$ , can be inferred by Equation (21). Seen from the numerator, all the retained blocks in the process of model selection meet  $\sum_{i=1}^{n} \zeta_{ik} > K_{ne}$ , then  $\sum_{k=1}^{K_{ne}} \sum_{i=1}^{n} \zeta_{ik} > \sum_{k=1}^{K_{ne}} K_{ne}$ , and  $K_{ne} < \sqrt{n}$  can be acquired, which can be seemed as the resolution limit of SSBM. Therefore, when there is no prior knowledge on signed networks, the heuristic information can provide a good choice for initializing the maximum detectable block number, i.e.,  $K_{max} = \sqrt{n}$ .

## **2.3.** Time Complexity Analysis

The pseudo code of SSBM learning algorithm is presented in Algorithm 1, and the flow of execution is visualized in Fig. 3. Obviously, the most time-consuming parts of SSBM learning algorithm are repeat loop and foreach loop. The foreach loop is responsible for evaluating block, selecting block, estimating parameters, and discarding block. The repeat loop repeats above calculations until cost function converges, and then the optimal model and latent variable *Z* are obtained (Lines 24 and 25). Specifically, in foreach loop,  $\zeta_{ik}^{(t)}$  is calculated first, and then  $\phi_k^{(t)}$  (Lines 7 to 9). Since  $\zeta_{ik}^{(t)}$  is posterior probability after the  $t^{th}$  iteration, the block quality can be evaluated based on  $\sum_{i=1}^{n} \zeta_{ik}^{(t)}$  and  $\frac{c}{2}$ . For example, if  $\sum_{i=1}^{n} \zeta_{ik}^{(t)} < \frac{c}{2}$ , i.e., block *k* cannot be supported by data,  $\phi_k^{(t)} = 0$  and block *k* will be discarded (Line 15); otherwise, block *k* will be selected (Line 10). Then the parameters of selected block will be estimated (Lines 11 and 12).

The time complexity of calculating  $\zeta_{.k}$  and  $\phi_k$  in Lines 7 and 8 are  $O(nK_{max})$  and  $\lambda_{.k}$  and  $u_{.k}$  in Lines 11 and 12 are  $O(n^2)$ , respectively. Thus, the time complexity of executing a complete foreach loop is  $O(n^2K_{max} + nK_{max}^2)$ . The time complexity of calculating  $C(N, g^{(t)})$  is  $O(n^2K_{max} + K_{max})$ . The above calculations cost  $O(2n^2K_{max} + nK_{max}^2 + K_{max})$  in total. Assuming that the algorithm converges after *T* iterations, the repeat loop is  $O(Tn^2K_{max})$ . Therefore, the time complexity of SSBM is  $O(Tn^2K_{max}(K_{max} - K_{min}))$  due to the while loop executing  $K_{max} - K_{min}$  times.

#### 3. Validation

In this section, the generalization, robustness and scalability of SSBM will be validated by comparisons with stateof-the-art methods on synthetic and real-world networks. The programs of all algorithms are developed using MATLAB 2010b and run on a computer with a 4-core CPU with a 3.20, 8GB RAM, and 64-bit Windows 10 operating system. For all the compared algorithms,  $K_{min} = 1$  and  $K_{max} = 10$  are set uniformly, and for SSBM,  $K_{min} = 1$  and  $K_{max} = \sqrt{n}$ are set in term of the aforementioned heuristic information, and  $\Phi$  is initialized by  $(1/K_{max}, ..., 1/K_{max})$ ,  $\Lambda$  is initialized randomly,  $\epsilon = 10^{-4}$ .

#### **3.1. Metrics and Baselines**

The accuracy of signed network structure discovery is evaluated by normalized mutual information (NMI) [34]. Assume that A and B are the real and discovered network structure partitions, respectively. The NMI can be calculated as:

$$NMI(A, B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}m_{ij}\log{(\frac{m_{ij}M}{m_i.m_{.j}})}}{\sum_{i=1}^{C_A}m_{i.}\log{(\frac{m_{i}}{M})} + \sum_{j=1}^{C_B}m_{.j}\log{(\frac{m_{.j}}{M})}}$$

where *M* is confusion matrix, and  $m_{ij}$  denotes the number of nodes belonging to block *i* of *A* and block *j* of *B* at the same time.  $C_A$  and  $C_B$  are the number of blocks in *A* and *B*, respectively.  $m_{ij}$  and  $m_{ij}$  represent the sum of the elements in Row *i* and Column *j* in *M*, respectively. If the discovered partition *B* is identical to the real partition *A*, NMI(A, B) =1; otherwise, NMI(A, B) = 0.

Five classical signed network structure methods, namely, VBS [4], SSL [18], SISN [3], FEC [14], and DM [11], are selected as compared methods.

Algorithm 1: SSBM Learning Algorithm **Input:**  $N, K_{min}, K_{max}$ **Output:**  $g_{best}, Z_{best}$ 1 Initialize  $\Phi^{(0)}; \Lambda^{(0)}; t \leftarrow 0; K_{ne} \leftarrow K_{max}; \varepsilon;$ 2  $u_{ik}^{(0)} \leftarrow \prod_{j=1}^{n} \prod_{h=1}^{3} (\lambda_{kjh}^{(0)})^{\delta(a_{ij},2-h)}$ , for i = 1, ..., nand  $k = 1, ..., K_{max}$ ;  $C_{min} \leftarrow +\infty; g^{(0)} = (K_{max}, \Phi^{(0)}, \Lambda^{(0)})$ 3 while  $K_{ne} \ge K_{min}$  do 4 repeat  $t \leftarrow t + 1;$ 5  $\begin{aligned} foreach \ k &= 1 \ to \ K_{max} \ do \\ \zeta_{ik}^{(t)} \leftarrow \frac{\phi_{k}^{(t-1)} u_{ik}^{(t-1)}}{\sum_{l=1}^{K_{max}} \phi_{l}^{(t-1)} u_{il}^{(t-1)}}, \ for \ i = 1, ..., n; \\ \phi_{k}^{(t)} \leftarrow \frac{max\{0, \sum_{i=1}^{n} \zeta_{ik}^{(t)} - \frac{c}{2}\}}{\sum_{l=1}^{K_{max}} max\{0, \sum_{i=1}^{n} \zeta_{il}^{(t)} - \frac{c}{2}\}}; \\ \phi_{q}^{(t)} \leftarrow \phi_{q}^{(t)} (\sum_{l=1}^{K_{max}} \phi_{l}^{(t)})^{-1}; \end{aligned}$ 6 7 8 q  $\begin{aligned} \varphi_{q} & (\varphi_{q} (\Sigma_{l=1} \ \varphi_{l}))^{*}, \\ \mathbf{if} \ \phi_{k}^{(t)} &> 0 \ \mathbf{then} \\ & \lambda_{kjh}^{(t)} \leftarrow \frac{\sum_{i=1}^{n} \zeta_{ik} \delta(a_{ij}, 2-h)}{\sum_{i=1}^{n} \zeta_{ik}}, \ \mathbf{for} \ j = 1, .., n, \\ & h = 1, 2, 3; \\ & u_{ik}^{(t)} \leftarrow \prod_{j=1}^{n} \prod_{h=1}^{3} (\lambda_{kjh}^{(t)})^{\delta(a_{ij}, 2-h)}, \ \mathbf{for} \\ & i = 1, .., n; \end{aligned}$ 10 11 12 end 13 if  $\phi_k^{(t)} \leq 0$  then  $| K_{ne} \leftarrow K_{ne} - 1;$ 14 15 end 16 end 17  $g^{(t)} = (K_{ne}, \Phi^{(t)}, \Lambda^{(t)});$ 18  $C(N, g^{(t)}) \leftarrow \frac{K_{ne}(c+1)}{2} \log n + \frac{c}{2} \sum_{\phi_k > 0} \log \phi_k + \frac{K_{ne}(c+1)}{2} (1 + \log \kappa_d) - \sum_{i=1}^n \log \sum_{k=1}^{K_{max}} \phi_k^{(t)} u_{ik}^{(t)};$ 19 20 until  $(C(N, g^{(t-1)}) - C(N, g^{(t)})) < \varepsilon$ 21 if  $C(N, g^{(t)}) < C_{min}$  then 22  $C_{min} \leftarrow C(N, g^{(t)});$   $g_{best} \leftarrow g^{(t)};$   $Z_{best} \sim multinomial(\zeta^{(t)});$ 23 24 25 end 26 27 end

- *VBS* [4] is a learning algorithm with model selection, which is proposed by extending the standard SBM to signed SBM in the framework of variational Bayesian.
- *SSL* [18] is a variational Bayesian EM algorithm based on approximate evidence, and a signed stochastic block model is defined by explicitly modeling the density and noise distribution of the edges.
- *SISN* [3] is a signed network community discovery method based on statistical reasoning, and a model selection strategy based on the minimum description length (MDL) is presented.
- FEC [14] is a method based on random walk model to

discover signed network communities by alternately executing FC (finding a municipality) and EC (extracting a community).

• *DM* [11] is a local search method based on equilibrium theory proposed by Doreian and Murvar, which detects networks structures by minimizing the noise of signed networks.

# 3.2. Validation on Synthetic Signed Networks

In this section, the effectiveness of SSBM will be validated using synthetic networks compared with five state-ofthe-art methods.

# 3.2.1. Synthetic Datasets Generation

1

For the sake of fairness, synthetic signed networks are mostly generated using the model proposed in [14]:

$$Model_{sign} = SG(c, m, k, p_{in}, p-, p+)$$
(24)

where *c*, *m*, and *k* are the number of blocks, the number of nodes in a block, and average node degree, respectively.  $p_{in}$  is a parameter for controlling cohesiveness. It represents the probability of generating edges between nodes within blocks. The closer the value is to 1, the more distinct the network structure. *p*- and *p*+ are parameters for controlling the noise. They represent the probabilities of generating negative edges within blocks and positive edges between blocks, respectively. The larger the two values, the more complex the network structure.

Five kinds of synthetic networks are generated using the above generative model.

- *Network I*: It is generated by parameter configuration  $(4,32,32,p_{in},0,0)$ , where  $p_{in}$  increases from 0.0 to 1.0, and the step size is 0.1. It is a balanced signed network where positive edges only exist within blocks while negative edges only exist between blocks.
- *Network II*: It is generated by parameter configuration (4,32,32,0.6,p-,0), where *p*- increases from 0.0 to 0.5, and the step size is 0.05. The negative edges within blocks can be seen as network noises. As *p*- increases, there are more negative edges within blocks.
- *Network III*: It is generated by parameter configuration (4,32,32,0.6,0,p+), where *p*+ increases from 0.0 to 0.5, and the step size is 0.05. The positive edges between blocks can be seen as noises. As *p*+ increases, there are more positive edges between blocks.
- *Network IV*: It is generated by parameter configuration (4, 32, 32, 0.6, p-, 0.5), where *p* also increases from 0.0 to 0.5, and the step size is also 0.05. There are two kinds of noises in the generated network, i.e., positive edges between blocks and negative edges within blocks, and it is more complex than network II.





Figure 3: The visual flow chart of SSBM.

• *Network V*: Similar to Network IV, it is generated by parameter configuration (4, 32, 32, 0.6, 0.5, p+), where negative edges within blocks are generated with the probability of 0.5, and positive edges between blocks are generated according to *p*+.

Networks II-V are unbalanced signed networks that are mainly used to validate the robustness of the proposed method. To further verify the ability of SSBM to discover multiple structures, Network VI containing community and bipartite structures is generated by the following method.

• *Network VI*: It is generated by dividing all nodes into two communities and two bipartites, and each block contains 32 nodes. The number of positive, negative, or null edges generated within and between blocks is subject to multinomial distribution with parameters of  $\pi_{kk}$  and  $\pi_{kq}$ , respectively. Specifically,  $\pi_{kk}$  and  $\pi_{kq}$  are set as follows:

$$\begin{split} \pi_{11} &= \{0.6, 0.1, 0.3\}, \pi_{12} = \{0.1, 0.2, 0.7\}, \pi_{13} = \\ \{0.1, 0.2, 0.7\}, \pi_{14} = \{0.1, 0.2, 0.7\}; \\ \pi_{22} &= \{0.2, 0.1, 0.7\}, \pi_{23} = \{0.01, 0.4, 0.59\}, \pi_{34} = \\ \{0.01, 0.4, 0.59\}; \\ \pi_{33} &= \{0.01, 0.01, 0.98\}, \pi_{34} = \{0.01, 0.4, 0.59\}; \\ \pi_{44} &= \{0.01, 0.01, 0.98\}. \end{split}$$

# 3.2.2. Results and Analysis

Fig. 4 presents the experimental results on synthetic networks. For balanced signed networks (Fig. 4 (a)), SSBM and VBS show excellent performance in community structure discovery. With an increase of  $p_{in}$  from 0 to 1, both of them can find communities accurately, indicating that they are insensitive to cohesiveness within blocks. The accuracy of SISN decreases slightly only when  $p_{in} = 0.5$ , and SSL and DM are worse than SISN. SSL is incapable of discovering community structure when  $p_{in} = 0$  since the cohesiveness in the community is very weak, while with an increase of  $p_{in}$ , the community discovery ability of SSL is gradually enhanced. In contrast to SSL, when cohesiveness in the community is very strong, e.g.,  $p_{in} = 1$ , DM cannot discover community structure. When  $p_{in} < 1$ , DM has a good ability of community discovery.

Experimental results on five kinds of unbalanced signed networks are presented in Fig. 4 (b)-(f), respectively. In Fig. 4 (b), SSBM, VBS, and SSL perform the best. They are not affected by noises within communities, and can accurately find community structure. SISN can detect the community structure in most cases with noise ratio p-varying, and it achieves the best performance (NMI = 0.923) when p-=0.15. For FEC and DM, the accuracy declines rapidly when p- increases gradually, indicating that they are more sensitive to noises within communities. In Fig. 4 (c), SSBM, VBS, and SSL are also the best ones, and they can accurately discover communities in all cases. As p+ increases, the performance of SISN and DM decreases slightly, but the NMIs are no less than 0.9 and 0.85, respectively. FEC is extremely sensitive to noises between communities, and it can accurately find the communities only when p + < 0.1; otherwise, it becomes ineffective.

As shown in Fig. 4 (d) and (e), SSBM and SSL still exhibit competitive performance. VBS can accurately discover community structure when  $p + \le 0.4$ , and it decreases slightly with an increase of p+. SISN is slightly worse than VBS, but the NMI is always larger than 0.83. FEC and DM are the worst. Specifically, for Network IV, DM performs better than FEC. FEC is invalid when  $p- \le 0.3$ . For Network V, FEC is better than DM. Experimental results on Network VI, which contains both community and bipartite structures, are shown in Fig. 4 (f). Specifically, the NMIs



Figure 4: Experimental results of six algorithms on six kinds of synthetic networks.

of SSBM, VBS, SSL, SISN, DM, and FEC are 1, 1, 0.958, 0.684, 0.957, and 0.273, respectively.

In summary, SSBM consistently exhibits the excellent performance for all the synthetic signed networks. Results on unbalanced signed networks further demonstrate that SSBM has good robustness to different network noise types and intensities, and shows strong generalization as well, i.e., it can accurately find multiple structures existed in the networks. Because real-world networks usually contain various noises and structures, SSBM has more advantages on discovering multiple structures in real-world networks than compared methods.

#### 3.3. Validation on Real-World Signed Networks

In this section, the ability of SSBM is further validated by dealing with real-world networks.

#### 3.3.1. Description of Real-World Datasets

In this experiment, we select three real-world datasets with ground truth, i.e., the Slovene Parliamentary Party Network (SPPN) [26], the Gahuku-Gama Subtribes Network (GGSN) [27], the Monastery Network (MN) [28], and a real-world dataset without ground truth, i.e., Country Network [29], to validate the effectiveness of SSBM.

 SPPN [26] is a signed network representing the relations among political parties in the Slovenian Parlia-

			-				
Networks	K <sub>true</sub>	SSBM	VBS	SSL	SISN	DM	FEC
SPPN	2	1/2	1/2	1/2	1/2	1/2	0.619/2
GGSN	3	1/3	1/3	1/3	0.528/ <b>3</b>	0.938/4	0.911/4
MN	3	1/3	0/1	0/1	1/3	0.86/ <b>3</b>	0.464/ <b>3</b>
Avg(rank)		1(1)	0.667(4)	0.667(4)	0.843(3)	0.933(2)	0.665(5)

Accuracy on three real-world signed networks

ment in 1994. The network contains 10 nodes, 2 communities, and 18 positive edges and 27 negative edges indicating that these political parties have similar or opposite political relations.

Table 1

- GGSN [27] describes political relationships between the Gahuku-Gama subtribes in 1954. The network contains 16 nodes, 3 communities, and 29 positive edges and 29 negative edges representing alliances or hostile political relations between political parties.
- MN [28] describes the emotion relations, i.e., like or dislike, among monks in the New England monastery. There are 18 nodes, 3 communities, and 51 positive edges and 58 negative edges.

#### 3.3.2. Results and Analysis

The results on three real-world networks with ground truth are shown in Table 1.  $K_{true}$  denotes the number of true communities, the value before "/" represents the NMI, and the value after "/" represents the number of communities discovered by algorithms. In the last line, the average NMI of each algorithm is calculated, and the value in "()" represents the ranking of each algorithm based on average NMI. SSBM outperforms all the compared methods because it can accurately discover network structures for three real-world networks. This further demonstrates that SSBM, using the block-to-node connection probability matrix  $\Lambda$  to reparameterize  $\Pi$  in the standard SBM, can capture more fine-grained network structure information and effectively improve the accuracy of multiple structure discovery for signed networks.

SSBM is further validated using a real-world dataset without ground truth, namely, the Country Network. It is derived from the Correlates of War dataset of countries from 1996 to 1999. After removing isolated nodes and sparsely connected components, the Country Network ultimately contains 144 nodes, 1099 positive and 144 negative edges, representing military alliances and military disputes, respectively. The partition result given by SSBM on the Country Network is shown in Fig. 5. Each color represents a cluster, and the solid dots represent nodes of the Country Network, and the positive and negative edges are denoted by solid and dotted gray lines, respectively. SSBM divides the Country Network into six clusters, where the number of nodes is 14 (yellow), 16 (blue), 18 (pink), 19 (white), 34 (green), and 43 (red), respectively. Obviously, this partition is quite reasonable: positive edges are mainly distributed within blocks, while negative edges mainly exist between blocks, and there are basically no controversial nodes.



**Figure 5:** The partition visualization of SSBM for Country Network.

The above validations show that SSBM has excellent accuracy and good generalization ability. It can detect the block structures in the signed networks reasonably and efficiently without any prior knowledge, and is more applicable for handling real-world exploratory signed networks.

## 3.4. Validation of Scalability

For validating scalability, a series of unbalanced synthetic signed networks with different scales are generated using the aforementioned model (Equation (24)). All of the unbalanced synthetic networks contain four clusters, i.e., c in Equation (24) is 4. The *m* and *k* are set as 50, 100, 200, 500, 1000, 2500, and 5000 in sequence. That is, the numbers of nodes n are 200, 400, 800, 2000, 4000, 10000, and 20000, respectively. For all networks, p - = 0.5 and p + = 0.5. When  $n \leq 10000, p_{in} = 0.8$ ; otherwise,  $p_{in} = 0.4$ . Experimental results on accuracy and running time are presented in Table 2 and Fig. 6 (a), respectively. As seen from Table 2, SSBM and SSL are the best ones in all cases because the NMIs of two methods are always 1. VBS is slightly worse than SSBM and SSL, and it can discover network structures accurately in most cases. SISN is only capable of dealing with small networks. DM and FEC are worst, and the results indicate that the two methods are unadaptable to the signed networks containing complex noises.

In Fig. 6 (a), although SSBM and SSL have the same excellent accuracy, SSL needs a significant amount of time.

Nodes	200	400	800	2000	4000	10000	20000
SSBM	1	1	1	1	1	1	1
VBS	1	1	1	1	0.857	0.857	1
SSL	1	1	1	1	1	1	1
SISN	1	1	_	_	_	_	_
DM	0.04	0.007	0.022	0.004	0.003	0.0006	-
FEC	0.01	0.08	0.026	0.012	0	0.003	-

 Table 2

 Accuracy on large-scale synthetic signed networks

For example, when n is 200, SSBM and SSL take 0.09 seconds and 3.0 seconds, respectively. As network scale increases, the advantage of SSBM on dealing with large-scale networks is highlighted. When n is 10000, the running times of SSBM and SSL are 120 seconds and 4467.8 seconds, respectively. That is, SSBM is 37 times more efficient than SSL. Furthermore, when n is 20000, the performance gap between the two methods is even greater. SSBM only takes 739.2 seconds to discover network structures accurately, while SSL requires 34463.3 seconds.

Moreover, the experiments on real-world signed networks are designed to further the scalability of SSBM. The WikiEditor is one of the few real-world signed networks with a slightly larger scale [30]. It contains 21535 nodes, 269251 positive edges and 79004 negative edges. The nodes represent users participating in editing Wikipedia pages from Jan 2013 to July 2014, who were classified into benign users and vandals. Each edit of any user can belong to either revert or no-revert category. The edges between nodes are built on co-edit relations, that is, if most of the co-edits for two users belong to the same category, there is a positive edge between the two users; otherwise, there is a negative edge. To verify the scalability, a set of datasets are generated by extracting from the WikiEditor in proportions as 10%, 20%, 30%, 50%, 80% and 100%, respectively. The running times of all the algorithms on the datasets are shown in Fig. 6 (b).

SISN has the worst performance, and even cannot deal with the dataset composed of 10% of the WikiEditor. Therefore, it is not exihibited in Fig. 6 (b). The second worst is FEC, and it will terminate due to lack of memory when the size of dataset is more than 20% of the WikiEditor. The reason is that the edges of the WikiEditor are relatively dense, and FEC requires more computing resources in the execution process. DM performs better than FEC, and can handle datasets with the size less than half of the WikiEditor, although it takes a long time. SSL performs better than DM. SSL can detect the larger datasets with the size no more than 80% of the wikiEditor, and presents more faster running speed. SSBM and VBS are the best algorithms. Both of them can handle all the datasets in a short time, and when the size of dataset is equal or greater than 50% of the WikiEditor, VBS takes less time than SSBM. For example, when the dataset is the WikiEditor itself, SSBM takes 1572.06 seconds, while VBS takes only 819.16 seconds. The result seems to indicate that VBS is more efficient than SSBM in dealing with large-scale real-world signed networks, but it

is really not so. This is because that SSBM and VBS are not fair in the settings of model search space. In term of the heuristic information discussed previously, the model search space of SSBM for all experiments is  $[1, \sqrt{n}]$ , while VBS is [1, 10]. Therefore, the larger *n* is , the larger model space SSBM need to search, and the more time it will take. To be fair, a new experiment on setting the model search space of VBS as same as SSBM, i.e., the model search space of VBS is also  $[1, \sqrt{n}]$ , is presented. The corresponding result is labeled by VBS-1 in Fig. 6 (b). Obviously, the running time of VBS is much more than SSBM under the same condition. For example, when the size of dataset is 50% of the WikiEditor, SSBM takes 312.09 seconds while VBS takes 148449.14 seconds. As the size of dataset increasing, VBS will be invalid.

The experimental results in term of scalability show that the learning mechanism, synchronously carrying out parameter estimation and model selection, can significantly improve SSBM learning efficiency This make SSBM have more advantages and potentials in dealing with large-scale exploratory signed networks, especially unbalanced signed networks containing multiple structures and various noises.

# 3.5. Further Discussion

In this section, the advantages of SSBM in terms of generalization, robustness, and scalability are summarized. SSBM is essentially an extension of the standard SBM. First, we use block-to-node connection probability matrix  $\Lambda$  to reparameterize block-to-block connection probability matrix  $\Pi$ . This enables the extended model to capture structure information of networks from a more fine-grained perspective, and makes it more expressive than the standard SBM. Second, we let  $\Lambda$  be subject to a multinomial distribution, and it can explicitly depict the probabilities of generating positive, negative, and null edges and further exploit the applications of the standard SBM from unsigned networks to signed networks. These two extensions enable SSBM to accurately discover multiple structures in various signed networks.

The SSBM has good robustness to noises because the parameter  $\Lambda$  can explicitly model noise densities in the signed networks. Taking  $\Lambda_{kj} = (\lambda_{kj1}, \lambda_{kj2}, \lambda_{kj3})$  as an example,  $\lambda_{kj1}$  and  $\lambda_{kj2}$  represent the probabilities that generating positive and negative edges between any node in block *k* and node *j*, respectively. If node *j* belongs to block *k*, then  $\lambda_{kj2}$  is the probability of generating a negative edge between any node in block *k* and node *j*, i.e., intrablock noise, and the ex-



Figure 6: Running time on synthetic and real-world large-scale signed networks.

pectation  $\sum_{j \in k} \sum_{i \in k} \lambda_{kj2}$  denotes the density of intrablock noises. Conversely, if node *j* does not belong to block *k*, then  $\lambda_{kj1}$  denotes the probability that a node in block *k* generates a positive edge to node *j*, and its expectation  $\sum_{j \notin k} \sum_{i \in k} \lambda_{kj1}$ denotes the density of interblock noises. Therefore, SSBM can effectively overcome the influence of network noises with different types, e.g., negative edges in communities or positive edges between communities, and different densities, e.g., various configurations of *p*+ and *p*-.

According to the above analysis, the potential applications of SSBM can be briefly summarized in the following four perspectives: 1) discovering and extracting multiple structures in heterogeneous signed networks; 2) discovering and extracting single or multiple structures in signed networks containing complex noises; 3) discovering and extracting single or multiple structures in large-scale signed networks; 4) discovering and extracting single or multiple structures in exploratory signed networks.

It has to be said that  $K_{max}$  is set as  $\sqrt{n}$  for SSBM according to the heuristic information, i.e., the block search space is  $[1, \sqrt{n}]$ . Therefore, when the number of blocks contained in a signed network is greater than  $\sqrt{n}$ , SSBM cannot correctly detect the network structures.  $K_{max} = \sqrt{n}$  is the resolution limit of SSBM. For this case, a strategy can be employed to flexibly set the block search space according to the space complexity  $O(\sqrt{n})$ , i.e., set  $[K_{min}, K_{max}]$  as  $[(m-1)\sqrt{n} + 1, m\sqrt{n}](m = 1, 2, ..., \sqrt{n})$ . This strategy can ensure that SSBM can accurately discover the network structures, effectively reduce the search space, and improve the algorithm's efficiency.

# 4. Related Work

This section introduces previous studies closely related to our method from discriminant and principled perspectives.

## 4.1. Discriminant Methods

Discriminant methods usually divide nodes to different clusters based on predefined optimization objectives [31] or heuristic information [32, 33]. In 1996, Doreian and Murvar proposed a frustration-based signed network community discovery method named as DM [11], which detect communities by minimizing network noises. There are two kinds of noises, i.e., negative edges in communities and positive edges between communities. Besides community structure, DM can also detect bipartite structure and coexistence of community and bipartite structures. In view of this, DM can be seen as the early multiple structure discovery approach. Bnasal et al. proposed a community detection method for signed networks by maximizing the sum of positive edges within and negative edges between communities [12]. Traag and Bruggeman proposed a modularity-based community partition method by maximizing the modularity of signed networks [13]. In addition, evolutionary computation and nonnegative matrix factorization methods are used to detect the community structure. For instance, in 2016, Li et al. proposed an optimization method based on a multiobjective particle swarm to discover communities [16]. In 2018, Zhu et al. used evolutionary algorithm to detect community structure of unbalanced signed networks [17]. Recently, Li et al presented a new form of non-negative matrix factorization and a probabilistic surrogate learning function for community detection, but this method can only be applicable to unsigned networks [35].

Most of discriminant methods mentioned above can only discover community structure of signed networks. Moreover, the scalability of these methods is very limited. In view of this, Yang et al. proposed a fast community discovery method FEC by employing a Markov stochastic process [14]. Anchuri et al. proposed a spectral method to find communities by optimizing modularity or other objective functions [15]. These methods can effectively handle signed networks with thousands of nodes, but they still cannot detect multiple structures and are very sensitive to network noises. For all of discriminant methods, there is a common problem that their performances overly rely on predefined optimization objectives or heuristic information.

# 4.2. Principled Methods

For principled methods, network structures can be discovered from a network generation perspective by fitting probabilistic model to observed networks [36, 37]. For example, Zhao et al. proposed a probabilistic model named as SISN for detecting community structure of signed networks [3]. Compared with discriminant methods, principled methods can find the intrinsic structures more accurately and have good interpretability, but they also cannot discover multiple structures contained in signed networks. For this proplem, Yang et al. proposed a signed stochastic block model and presented a variational Bayesian learning method SSL for parameter estimation and model selection [18]. SSL can effectively discover community or bipartite structure, but it fails to detect coexistence of them. Zhao et al. proposed a mathematically principled method, namely VBS, to discover the multiple structures in signed networks [4]. This method firstly presented a probabilistic model to characterize the signed networks with community, bipartite or coexistence of them, and then deduced the approximate distribution of model parameters by utilizing a variational Bayesian approach. VBS is the classical method of multiple structure discovery for signed networks.

In recent years, some scholars carried out the researches on multiple structure discovery for unsigned networks. Liu et al. proposed a generative node-attribute network model, namely GNAN, by combining topological information of network and attribute information of nodes[38]. GNAN can detect communities more accurately due to utilizing node attributes, and detect multiple structures including bipartite, core-periphery, and their mixing structure. He et al. developed a Bayesian probabilistic mixture model NEGCD by incorporating network embedding into topological structure of network[40]. NEGCD can detect assortative structure, i.e., community, and disassortative structure, i.e., bipartite, and mixing structure. By contrast, the researches on detecting multiple structures in signed networks are quite few, because it is more difficulty and challenging but worth deeply studying.

Moreover, high time complexity is a common problem faced by all principled methods, making them fail to deal with large-scale networks. To this end, some scholars conducted the researches on scalability. Li et al introduced a new belief-dynamic-based Markov clustering technique, called BMCL, for large-scale network community detection, but BMCL just can only deal with unsigned networks [39]. Zhao et al. proposed a block-wise SBM learning algorithm named as BLOS to improve the scalability of current SBM-based learning methods [41]. Different from existing methods, BLOS can implement model selection and parameter estimation simultaneously by introducing the minimum message length (MML) criterion into a block-wise EM algorithm. BLOS has good performance on dealing with large-scale networks in real applications. On this basis, Li et al. proposed a reparameterized SBM algorithm RSBM, which is suitable for unsigned networks with heterogeneous distributions of node degree and block size [42]. However, the above methods are

merely designed to focus on structural features of unsigned networks, and can apply to signed networks.

# 5. Conclusions

In this paper, we propose a novel reparameterized signed stochastic block model SSBM to characterize multiple structures in signed networks and present a scalable learning algorithm with model selection ability by integrating MML and CEM. The generalization, robustness, and scalability of SSBM are validated on synthetic and real-world networks. Experimental results demonstrated the superiority of SSBM by comparing it with five representative network structure discovery methods. Future works will be studied from two aspects: the first is studying SBM variations for various networks based on reparameterization, e.g., heterogeneous networks, overlapping networks, multilayer networks, and dynamic networks; the second is presenting a general framework of SBM physical parallel learning, which can further improve scalability in dealing with large-scale networks.

# Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under grants Nos.61902145, 62172185, 61876069, 61976102, and U19A2065; the National Key R&D Program of China under grants Nos. 2021ZD0112501 and 2021ZD0112502; the International Cooperation Project under grant No. 20220402009GH; and Jilin Province Natural Science Foundation under Grant No. 20200201036JC.

# References

- J. Chen, H. Wang, L. Wang, and W. Liu, "A dynamic evolutionary clustering perspective: community detection in signed networks by reconstructing neighbor sets," Physica A: Statistical Mechanics and its Applications, vol. 447, pp. 482-492, 2016.
- [2] S. Wang, M. Gong, H. Du, L. Ma, Q. Miao, and W. Du, "Optimizing dynamical changes of structural balance in signed networks based on memetic algorithm," Social Networks, vol. 44, pp. 64-73, 2016.
- [3] X. Zhao, B. Yang, X. Liu and H. Chen, "Statistical inference for community detection in signed networks," Physical Review E, vol. 95, pp. 42313-42320, 2017.
- [4] X. Zhao, X. Liu, and H. Chen, "Network modelling and variational Bayesian inference for structure analysis of signed networks," Applied Mathematical Modelling, vol. 61, pp. 237-254, 2018.
- [5] X. Liu, B. Yang, H. Chen, K. Musial, H. Chen, Y. Li, W. Zuo, "A Scalable Redefined Stochastic Blockmodel," ACM Transactions on Knowledge Discovery from Data, vol. 15(3), pp. 1-28, 2021.
- [6] C. He, Q. Zhang, Y. Tang, S. Liu, and J. Zheng, "Community detection method based on robust semi-supervised nonnegative matrix factorization," Physica A: Statistical Mechanics and its Applications, vol. 523, pp. 279-291, 2019.
- [7] S. Rahimi, A. Abdollahpouri, and P. Moradi, "A multi-objective particle swarm optimization algorithm for community detection in complex networks," Swarm and Evolutionary Computation, vol. 39, pp. 297-309, 2018.
- [8] B. Liu, H. Su, L. Wu, et al. Controllability for multi-agent systems with matrix-weight-based signed network. Applied Mathematics and Computation, vol. 411, pp. 126520-126534, 2021.
- [9] J. Tang, Y. Chang, C. Aggarwal, and H. Liu, "A survey of signed network mining in social media," Acm Computing Surveys, vol. 49, pp. 1-37, 2016.

- [10] M.E.J. Newman, "Communities, modules and large-scale structure in networks," Nature. Physics., vol. 8, pp. 25-31, 2012.
- [11] P. Doreian, and A. Mrvar, "A partitioning approach to structural balance," Social Networks, vol. 18, pp. 149-168, 1996.
- [12] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," Machine Learning, vol. 56, pp. 89-113, 2014.
- [13] V.A. Traag, and J. Bruggeman, "Community detection in networks with positive and negative links," Physical Review E, vol. 80, pp. 36115-36120, 2009.
- [14] B. Yang, W. Cheung, and J. Liu, "Community mining from signed social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 19, pp. 1333-1348, 2007.
- [15] P. Anchuri, and M. Magdon-Ismail, "Communities and balance in signed networks: a spectral approach," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 235-242, 2012.
- [16] Z. Li, L. He, and Y. Li, "A novel multiobjective particle swarm optimization algorithm for signed network community detection," Appled Intelligence, vol. 44, pp. 621-633, 2016.
- [17] X. Zhu, Y. Ma, and Z. Liu, "A novel evolutionary algorithm on communities detection in signed networks," Physica A: Statistical Mechanics and its Applications, vol. 503, pp. 938-946, 2018.
- [18] B. Yang, X. Liu, Y. Li, and X. Zhao, "Stochastic blockmodeling and variational Bayes learning for signed network analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 29, pp. 2026-2039, 2017.
- [19] C.S. Wallace, and D.M. Boulton, "An information measure for classification," The Computer Journal, vol. 11, pp. 185-194, 1968.
- [20] G. Celeux, S. Chrtien, F. Forbes, and A. Mkhadri, "A component-wise EM algorithm for mixtures," Journal of Computational and Graphical Statistics, vol. 10, pp. 697-712, 2001.
- [21] T.A.B. Snijders, K. Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," Journal of Classification, vol. 14(1), pp. 75-100, 1997.
- [22] A.D. Lanterman, "Schwarz, Wallace, and Rissanen: Interwining themes in theories of model selection," International Statistical Review, vol. 69(2), pp. 185-212, 2001.
- [23] M.A.T. Figueiredo, A.K. Jain, "Unsupervised learning of finite mixture models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24(3), pp. 381–396, 2002.
- [24] D.M. Titterington, A.F.M. Smith, U.E. Makov, "Statistical analysis of finite mixture distributions," Chichester, U.K.: John Wiley & Sons, 1985.
- [25] K.P. Murphy, "Machine Learning: A Probabilistic Perspective," MIT Press, 2012.
- [26] S. Kropivnik, and A. Mrvar, "Ananalysis of the Slovene parliamentary parties network," Developments in Statistics and Methodology, vol. 12, pp. 209-216, 1996.
- [27] K.E. Read, "Cultures of the central highlands, newguinea, Southwest," Anthropol, vol. 10, pp. 1-43, 1954.
- [28] S.F. Sampson, "Crisis in a cloister," Ph.D. dissertation, Ph.D. Thesis, Cornell University, Ithaca, 1969.
- [29] P. Doreian, and A. Mrvar, "Structural balance and signed international relations," Society Structure, vol. 16, pp. 1-49, 2015.
- [30] S. Kumar, F. Spezzano, V.S. Subrahmanian, "Vews: A Wikipedia Vandal Early Warning System," In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.
- [31] X. Zhu, Y. Ma, Z. Liu, "A novel evolutionary algorithm on communities detection in signed networks." Physica A: Statistical Mechanics and its Applications, vol. 503, pp. 938-946, 2018.
- [32] P. Esmailian, M. Jalili, "Community Detection in Signed Networks: the Role of Negative ties in Different Scales," Scientific Reports, vol. 5, pp. 14339-14355, 2015.
- [33] J. Chen, H. Wang, L. Wang, "A dynamic evolutionary clustering perspective: Community detection in signed networks by reconstructing neighbor sets," Physica A Statistical Mechanics & Its Applications, vol. 447, pp. 482-492, 2016.

- [34] L.I. Kuncheva, S.T. Hadjitodorov Wang, "Using diversity in cluster ensembles,"In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1214-1219, 2004.
- [35] H. Li, L. Wang, Y. Zhang, M. Perc, "Optimization of identifiability for efficient community detection," New Journal of Physics, vol. 22(6), pp. 63035-63044, 2020.
- [36] P. J. Mucha, T. Richardson, K. Macon, "Community structure in time-dependent, multiscale, and multiplex networks," science, vol. 328(5980), pp. 876-878, 2010.
- [37] P. Doreian, "A multiple indicator approach to blockmodeling signed networks," Social Networks, vol. 30(3), pp. 247-258, 2008.
- [38] W. Liu, Z. Chang, C. Jia, Y. Zheng, "A generative node-attribute network model for detecting generalized structure and semantics," Physica A, vol. 588, pp. 126557-126570, 2021.
- [39] H. Li, W. Xu, C. Qiu, J. Pei, "Fast Markov clustering algorithm based on belief dynamics", IEEE Transactions on Cybernetics, doi: 10.1109/TCYB.2022.3141598, 2022.
- [40] D. He, Y. Wang, J. Cao, W. Ding, S. Chen, Z. Feng, B. Wang, Y. Huang, "A network embedding-enhanced Bayesian model for generalized community detection in complex networks," Information Sciences, vol. 575, pp. 306-322, 2021.
- [41] B. Yang, X. Zhao, "On the scalable learning of stochastic blockmodel," In Proceedings of the Twenty-ninth Aaai Conference on Artificial Intelligence, pp. 360-366, 2015.
- [42] Y. Li, H. Chen, and B. Yang, "Reparameterized stochastic block model adaptive to heterogeneous degree and block distributions." IEEE Access, vol. 6, pp. 37615-37626, 2018.