Applied Mathematical Modelling 61 (2018) 237–254

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Network modelling and variational Bayesian inference for structure analysis of signed networks

Xuehua Zhao[a], Xueyan Liu[b], Huiling Chen[c,*]

[a] School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China
[b] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[c] College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou 325035, China

ARTICLE INFO

Article history:
Received 8 July 2017
Revised 20 April 2018
Accepted 23 April 2018
Available online 28 April 2018

Keywords:
Network modelling
Variational Bayesian
Structure mining
Signed network

ABSTRACT

Currently, structure analysis of signed networks with positive and negative links has received wide attention and is becoming a research focus in the area of network science. In recent years, many community detection methods for signed networks have been proposed to analyze the structure of signed networks. However, current methods can only efficiently analyze the signed networks with the single community structure and unable to analyze the signed networks with the coexisting structure of communities and peripheral nodes, bipartite, or other structures. To address this problem, in this study, we present a mathematically principled method for the structure analysis of signed networks with positive and negative links, in which a probabilistic model firstly is proposed to model the signed networks with the single community or the coexisting structure, and a variational Bayesian approach is deduced to learn the approximate distribution of model parameters. For determining the optimal model, we also deduce a model selection criterion based on the evidence theory. In addition, to efficiently analyze the large signed networks, we propose a fast learning version of our algorithm with the time complexity $O(k^2E)$ where k is the number of groups and E is the number of links. In our experiments, the proposed method is validated in the synthetic and real-world signed networks, and is compared with the state-of-the-art methods. The experimental results demonstrate that the proposed method can more efficiently and accurately analyze to the structure of signed networks than the state-of-the-art methods.

1. Introduction

Signed networks consist of the nodes, positive links and negative links, in which the nodes represent the individuals, the positive links represent like, trust or support relationship and the negative links represent dislike, distrust or oppose relationship [1,2]. In contrast to unsigned networks only describing whether the relationship between two individuals exists or not [3–5], signed networks may contain more information by extending the single relationship to the positive and negative relationships. As such, signed networks with positive and negative links are receiving wide attention in recent years [6–8]. The signed networks usually fall into two categories according to whether the link has a direction or not: undirected signed networks and directed signed networks. For undirected signed networks, structure balance theory is an important theory to analyze the undirected signed networks, in which a triad is balanced if the product of the signs in the triad is positive

* Corresponding author.
E-mail addresses: lcrlc@sina.com (X. Zhao), dyyzlxy@163.com (X. Liu), chenhuiling.jlu@gmail.com (H. Chen).

https://doi.org/10.1016/j.apm.2018.04.013
0307-904X/

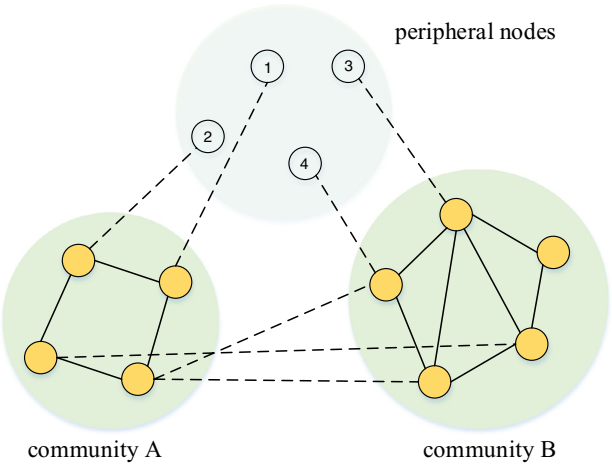

Fig. 1. Example of the mixed structure in the signed network. The circles represents nodes, the number in the nodes is the label of nodes, the dotted lines and solid lines respectively represent the negative links and positive links.

and it is imbalanced if the product of the signs is negative [9]. However balance theory only can analyze the undirected signed networks. For analyzing the directed signed networks, Leskovec et al. proposed the status theory, in which a positive directed link indicates that the creator of the link views the recipient as having higher status, and a negative directed link indicates that the recipient is viewed as having lower status [10]. These relative levels of status can be propagated along multi-step paths of signed links, often leading to different predictions than balance theory.

Structure analysis is an important problem in the network studies since network structures are closely related to the functions and evolution of systems. Community structure, which is the dense subnetwork within a larger network, is the best-studied structure in networks. In general, the links in communities are dense but the links between communities are sparse [11]. Because there are negative links in the signed networks, the communities in signed networks also show another characteristic that is most of the links in communities are the positive links and most of the links between communities are the negative links. In this sense, the communities in the signed network are consistent with the clusters defined in balance theory in the social science [9,12], where a strongly (or weakly) balanced network can be divided into two (or k) clusters, so that all the links within the clusters are positive links and all the links between the clusters are negative links. To analyze the signed networks, until now, many methods have been proposed to find the communities in signed networks. The representative methods mainly fall into two categories.

One class of methods is the discriminative method which usually divides the nodes into different communities based on either predefined optimization objectives (such as modularity) or heuristics (such as a random walk model). For example, Doreian and Mrvar proposed a frustration-based method (referred to as DM), which finds the communities by minimizing the sum of the number of negative links within communities and positive links between communities [13]. The method proposed by Bansal et al. finds the communities by maximizing the number of positive intra-cluster links and negative inter-cluster links or minimizing the number of negative intra-cluster links and positive inter-cluster links [14]. Traag and Bruggeman proposed a modularity-optimization-based algorithm for signed networks [15]. Yang et al. proposed a fast method based on Markov stochastic process (referred to as FEC) [16]. Anchuri and Magdon-Ismail proposed a generalized spectral method for signed network partition [17]. In addition, the multi-objective evolutionary methods also have been applied to signed network [18,19]. For this class of methods, their performances are closely related to the quality of the predefined objectives or heuristics. This is just their drawback since it is very difficult to design a good objective. Another class of methods is the model-based methods which find the communities in the signed networks by fitting the probability model to the signed networks. For example, Zhao et al. proposed an EM-based community detection method for the signed networks [2]. Yang et al. proposed a signed stochastic block model and its variational Bayes learning algorithm for signed networks [1].

The aforementioned methods mainly focus on the single community structure, however, for the real-world networks, in general, they could not only contain the single community structure, but the mixed structure of community and other structure such as peripheral nodes, bipartite or multipartite and so on. For example, Fig. 1 shows a mixed coexisting structure of community and peripheral nodes. For these peripheral nodes, if they are in the unsigned network, dividing the nodes 1 and 2 into the community A, the nodes 3 and 4 into the community B is reasonable, but in the signed network, dividing the nodes 1, 2, 3 and 4 in the same group is more reasonable since the links between these nodes and the nodes in the community A and B are negative. For these cases, current community detection methods for the signed networks usually directly merge the nodes of non-community structure into other communities so that they only can provide us with the imprecise results.

Block modelling is a form of statistical inference for the networks. The idea of block modelling for the network analysis is to find the structure of networks by fitting a specific block model to a network . For example, if we expect to analyze

the community structure, then we should first build a network model containing the community structure, then fit it to an observed network and analyze the community structure by the learned value of parameters of the model. Based on this idea, many methods are proposed to solve the problem of the coexisting structure analysis. The representatives of them are as follows. Daudin et al. proposed a mixture model which may model the community, star, bipartite or multipartite structure, and a variational EM learning algorithm to estimate the model parameters [20]. Newman and Leicht proposed a network model and adopted expectation maximization (EM) to estimate the parameters. Their method may be used to find the assortative and disassortative structures (communities and multipartite) [21]. Based on conventional stochastic block model (SBM), Latouche et al. proposed a variational Bayes EM algorithm for SBM and a model selection criterion based on a non-asymptotic approximation of the marginal likelihood [22]. The above algorithms may find the coexisting structure of community and other structures in the networks, but they only can be applied to the unsigned networks. To solve the signed networks, Yang et al. proposed a probabilistic model for signed networks in which the multinomial distribution is used to describe the distribution of the positive, negative and no links in the signed networks, and used the EM to learn the parameters [1]. Yang et al. proposed a signed stochastic block model and variational Bayes learning algorithm for signed networks [1]. However, these methods are mainly used to find the single structure (community or multipartite).

To address the above problems, inspired by block modelling, in this paper, we present a mathematically principled method for the structure analysis of the signed networks with positive and negative links, in which a probabilistic model firstly is proposed to model the signed networks, then a variational Bayesian approach is deduced to learn the approximate distribution of model parameters and variable. For determining the optimal model, we also deduce a model selection criterion based on evidence theory. In addition, to efficiently analyze the large signed networks, a fast learning version of our algorithm is presented, compared to the original algorithm, its time complexity is reduced from $O(k^2n^2)$ to $O(k^2E)$, where k, n, E are the number of groups, nodes and links, respectively. Since the large real-world signed networks usually are sparse, in this case, $E \ll n^2$, the proposed algorithm can efficiently analyze the large networks with more than 10^5 nodes. We validate the proposed algorithm in the synthetic and real-world signed networks and make comparisons with current algorithms. Our experimental results show the proposed method not only can efficiently analyze the single community structure but also can efficiently analyze the coexisting structure of community and other structure, and has more excellent performance than current algorithms.

The rest of this paper is organized as follows. In Section 2, we describe the proposed network model for exploring signed network structure, and provide the inference for parameter estimation. We also present a new model selection criterion and a fast version of our algorithm. In Section 3, we test the proposed algorithm in the synthetic and real-world signed networks. Finally, the discussions and conclusions are presented in Section 4.

2. Model and method

The proposed method, namely VBS, mainly includes two keys, which are network model and its learning algorithm. For network model, we present a new probabilistic model which can efficiently model the signed networks with the coexisting structure. For its learning algorithm, in the variational Bayesian framework, we deduce the approximate distribution of model parameters and the latent variable. Accordingly, the structure in signed networks can be inferred by analyzing the learned approximate distribution. When we solve the real-world networks without the prior structure information, we need the learning algorithm with the ability of model selection. The aim of model selection is determining the optimal model, which can make us solve the signed networks without any prior structure knowledge. For our model, determining the optimal model is determining the number of groups. In this study, we deduce a model selection criterion based on evidence theory. The VBS can work well for both directed and undirected signed networks, here, we introduce the VBS by the undirected case in detail.

2.1. Model

Let $\boldsymbol{a}$ denote the adjacency matrix of the signed network N containing n nodes. The element a_{ij} is equal to 1, −1 or 0 if there is a positive, negative or no link between the node i and the node j. Suppose all of the nodes are divided into K groups and the nodes in the same group have the similar connection pattern with the nodes of other groups. The proposed model is defined as follows

$$X = (K, \boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{\pi}), \tag{1}$$

where K is the number of groups. $\boldsymbol{\omega}$ is a K-dimension vector in which the element ω_k denotes the probability that a node is assigned to the group k, and $\sum_{k=1}^{K} \omega_k = 1$. $\boldsymbol{\pi}$ is a $K \times K \times 3$ matrix, where π_{lq1}, π_{lq2} and π_{lq3} denote the probability that there is a positive link, no link or negative link between a pair of nodes in the group l and q, respectively. In addition, the proposed model contains an indicating variable (or latent variable) $\boldsymbol{z}$, which is the $n \times K$ matrix containing the group information of nodes. $z_{ik} = 1$ if the node i is assigned to the group k, otherwise $z_{ik} = 0$.

Given the parameter $\boldsymbol{\omega}$, the probability distribution of $\boldsymbol{z}$ is as follows

$$p(\boldsymbol{z}|\boldsymbol{\omega}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \omega_k^{z_{ik}} \tag{2}$$

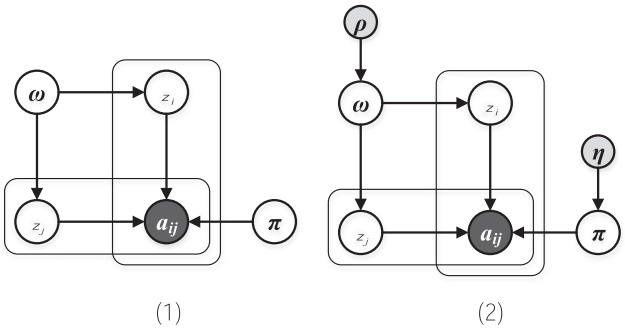

Fig. 2. Graphical representation of the proposed model. (1) The proposed model, (2) the proposed model in the Bayesian framework.

Given $\boldsymbol{z}$, a_{ij} follows the following multinomial distribution with parameter $\boldsymbol{\pi}$:

$$p(a_{ij}|\boldsymbol{z},\boldsymbol{\pi}) = \prod_{l,q=1}^{K}\prod_{h=1}^{3}\pi_{lqh}^{z_{il}z_{jq}\delta(a_{ij},2-h)}, \tag{3}$$

where $\delta(x, y)$ is the Kronecker function, if $x = y$, the function value is 1, otherwise the value is zero.

According to our model, one can generate a synthetic signed network by the following steps:

(1) Assign every node to its group according to the multinomial distribution with parameter $\boldsymbol{\omega}$, and build the variable $\boldsymbol{z}$.

(2) Generate the positive link, no link or negative link between two nodes according to the multinomial distribution with parameter $\boldsymbol{\pi}$ and the variable $\boldsymbol{z}$.

Accordingly, the complete log-likelihood can be written as follows

$$\log p(\boldsymbol{a},\boldsymbol{z}|K) = \sum_{i=1}^{n}\sum_{q=1}^{K} z_{iq}\log\omega_q + \sum_{i<j}^{n}\sum_{q,l}^{K}\sum_{h=1}^{3} z_{iq}z_{jl}\delta(a_{ij},2-h)\log\pi_{lqh} \tag{4}$$

When the priors of the model parameters $(\boldsymbol{\pi},\boldsymbol{\omega})$ are specified, we can describe the proposed model in a full Bayesian framework. Since $p(z_i|\boldsymbol{\omega})$ and $p(a_{ij}|\boldsymbol{z},\boldsymbol{\pi})$ satisfy the multinomial distribution, respectively, we can select the Dirichlet distribution as their conjugate prior distributions, as follows

$$p(\boldsymbol{\omega}|\boldsymbol{\rho}^0 = \{\rho_1^0,\ldots,\rho_K^0\}) = \mathrm{Dir}(\boldsymbol{\omega};\boldsymbol{\rho}^0) \tag{5}$$

$$p(\pi_{lq}|\boldsymbol{\eta}_{lqh}^0 = \{\eta_{lq1}^0,\eta_{lq2}^0,\eta_{lq3}^0\}) = \mathrm{Dir}(\pi_{lq};\boldsymbol{\eta}_{lqh}^0), \tag{6}$$

where $\boldsymbol{\rho}_q^0$ and $\boldsymbol{\eta}_{lqh}^0$ are the hyperparameters, which can be interpreted as the effective pseudo-occupations of respective groups in the prior, pseudo-observations of three types of links (positive, no and negative links) within or between the groups in the prior, respectively. In the full Bayesian framework, the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$ can be regarded as the random variables which follow the distributions with their respective hyperparameters. Fig. 2 shows the graphical representation of the proposed model.

The proposed model is a flexible model which not only can model various single structures in the signed networks, such as community, bipartite, multipartite and so on, but also can model the mixed coexisting structures of above single structures. This can be easily achieved only by simply setting the value of the parameter $\boldsymbol{\pi}$. It is also the reason that we can efficiently find the single structures or coexisting structures by fitting the proposed model to the observed networks.

2.2. Method

After the network model is given, we need to fit the model to the observed networks. In other words, to analyze the structure of network, we need to learn the parameters of the model, then analyze the network based on the learned values of parameters. Usually, the EM algorithm is a good learning method which estimates the values of parameters by maximizing the lower bound of log-likelihood $\mathcal{L}(N)$, however they cannot be directly used for our model since the posterior distribution of $\boldsymbol{z}$, under the condition of data and model parameters, cannot be explicitly derived as an input required by the EM. More specifically, z_i is correlated to others, that means the computation of its posterior distribution $P(z_i|N,\boldsymbol{\pi},\boldsymbol{\omega})$ is recursively dependent on the distribution of z_j for any $j \neq i$. Consequently, we adopt the variational Bayesian approach [23,24] to learn the approximate distributions of parameters and variable. Another advantage of the variational Bayesian approach is, in the Bayesian framework, we can calculate the lower bound of the evidence (marginal log-likelihood), and use the lower bound as an approximation of the evidence for model selection.

The log-likelihood $\mathcal{L}(N)$ of the network N (or the marginal log-likelihood of complete data) can be decomposed into two terms

$$\mathcal{L}(N) = \mathcal{L}(q(\cdot)) + KL(q(\cdot)||p(\cdot|N)), \tag{7}$$

where

$$\mathcal{L}(q(\cdot)) = \sum_{\boldsymbol{z}} \int\int q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}) \times \log\left\{\frac{p(N,\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})}{q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})}\right\} d_{\boldsymbol{\pi}}\, d_{\boldsymbol{\omega}} \tag{8}$$

$$KL(q(\cdot)||p(\cdot|N)) = -\sum_{\boldsymbol{z}} \int\int q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}) \times \log\left\{\frac{p(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}|N)}{q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})}\right\} d_{\boldsymbol{\pi}}\, d_{\boldsymbol{\omega}} \tag{9}$$

In Eqs. (7) and (9), $KL(q\|p)$ denotes the Kullback–Leibler divergence between the two distributions of $q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})$ and $p(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}|N)$. $p(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}|N)$ is the true posterior distribution of the variables $\boldsymbol{z}$ and the parameters $(\boldsymbol{\pi},\boldsymbol{\omega})$ given the network N, $q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})$ is an approximation of the true posterior distribution. In Eq. (8), $\mathcal{L}(q(\cdot))$ is called the lower bound of $\mathcal{L}(N)$. The Kullback–Leibler vergence satisfies $KL(q\|p) \geq 0$, with equality if, and only if, $q(\cdot) = p(\cdot)$. That means the lower bound $\mathcal{L}(q(\cdot))$ is equal to the log-likelihood $\mathcal{L}(N)$ when the vergence $KL(q\|p)$ vanishes. As a result, minimizing Eq. (9) with respect to $q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})$ is equivalent to maximizing the lower bound Eq. (8).

The variational approach aims at optimizing a lower bound of $\mathcal{L}(N)$ by approximating the true distributions of the parameters and variable. To obtain a computationally tractable algorithm, we use mean field approximation, one of the most popular forms of variational inference, in which we assume the posterior $q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})$ is a fully factorized approximation, which is written as follows

$$q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}) = q(\boldsymbol{\pi})q(\boldsymbol{\omega})\prod_{i=1}^{n} q(z_i), \tag{10}$$

where $q(z_i)$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\omega})$ denote the distributions of variables z_i, $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$, respectively.

Next, we need to seek the distributions of $q(z_i)$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\omega})$, which make the lower bound $\mathcal{L}(q(\cdot))$ largest. This requires us to deduce the expressions of the distributions $q(z_i)$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\omega})$, then we optimize the lower bound by iterating to update each expression in turn.

Proposition 1. *Given the distributions $q(\boldsymbol{\omega})$ and $q(\boldsymbol{\pi})$ of parameters $\boldsymbol{\omega}$ and $\boldsymbol{\pi}$, the optimal distribution $q(z_i)$ of z_i is the following multinomial distribution*

$$q(z_i) = M(z_i; 1, \tau_{i1}, \ldots, \tau_{iK}), \tag{11}$$

where τ_{ik} is the probability of node i belonging to group k, and satisfies:

$$\tau_{il} \propto e^{\psi(\rho_l) - \psi\left(\sum_{l=1}^{K}\rho_l\right)} \times \prod_{j\neq i}^{n}\prod_{q=1}^{K}\left(e^{\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij},2-h)\left(\psi(\eta_{lqh}) - \psi\left(\sum_{h=1}^{3}\eta_{lqh}\right)\right)}\right), \tag{12}$$

where $\psi(\cdot)$ is digamma function.

Proof. According to variational Bayes, the optimal distribution $q(z_i)$ of z_i can be derived as follows

$$\begin{aligned}
&\log q(z_i)\\
&= \mathrm{E}_{z^{\setminus i},\boldsymbol{\pi},\boldsymbol{\omega}}\left[\log p(N,\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})\right] + \text{const}\\
&= \mathrm{E}_{z^{\setminus i},\boldsymbol{\pi}}\left[\log p(N|\boldsymbol{z},\boldsymbol{\pi})\right] + \mathrm{E}_{z^{\setminus i},\boldsymbol{\omega}}\left[\log p(\boldsymbol{z}|\boldsymbol{\omega})\right] + \text{const}\\
&= \mathrm{E}_{z^{\setminus i},\boldsymbol{\pi}}\left[\sum_{i'<j}^{n}\sum_{l,q}^{K} z_{i'l}z_{jq}\sum_{h=1}^{3}\delta(a_{i'j},2-h)\log\pi_{lqh}\right] + \mathrm{E}_{z^{\setminus i},\boldsymbol{\omega}}\left[\sum_{i'=1}^{n}\sum_{l=1}^{K} z_{i'l}\log\omega_l\right] + \text{const}\\
&= \sum_{j\neq i}^{n}\Big(\sum_{l=1}^{K}\sum_{q=1}^{K} z_{il}\tau_{jq}\sum_{h=1}^{3}\big(\delta(a_{ij},2-h)\mathrm{E}_{\boldsymbol{\pi}}\left[\log\pi_{lqh}\right] + \sum_{q=1}^{K} z_{iq}\mathrm{E}_{\boldsymbol{\omega}}\left[\log\omega_q\right] + \text{const}\\
&= \sum_{l=1}^{K} z_{il}\left(\sum_{j\neq i}^{n}\sum_{q=1}^{K}\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij},2-h)\left(\psi\left(\eta_{lqh}\right) - \psi\left(\sum_{h=1}^{3}\eta_{lqh}\right)\right) + \left(\psi(\rho_q) - \psi\left(\sum_{k=1}^{K}\rho_k\right)\right)\right) + \text{const},
\end{aligned} \tag{13}$$

where $z^{\setminus i}$ denotes $\boldsymbol{z}$ of all the nodes except the node i. When $y \sim \mathrm{Dir}(y; a_1, a_2, \ldots, a_K)$, $E_y[\log(y_k)] = \psi(a_k) - \psi(\sum a_q)$, where $q \in \{1, 2, \ldots, K\}$. To simplify the calculations, the terms that do not depend on Z_i have been absorbed into the constant.

After we take the exponential of Eq. (13) and normalize it, the optimal distribution $q(z_i)$ in Eq. (11) of z_i can be obtained. □

Proposition 2. *Given the distribution $q(\boldsymbol{z})$, the optimal distribution $q(\boldsymbol{\omega})$ of the parameter $\boldsymbol{\omega}$ is the following Dirichlet distribution, which is the same form as its prior $p(\boldsymbol{\omega})$*

$$q(\boldsymbol{\omega}) = \mathrm{Dir}(\boldsymbol{\omega}; \boldsymbol{\rho}), \qquad \rho_q = \rho_q^0 + \sum_{i=1}^{n} \tau_{iq} \tag{14}$$

Proof. According to variational Bayes, the optimal distribution $q(\boldsymbol{\omega})$ can be derived as follows

$$\begin{aligned}
&\log q(\boldsymbol{\omega}) \\
&= \mathrm{E}_{\boldsymbol{z},\boldsymbol{\pi}}\big[\log p(N, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\omega})\big] + \text{const} \\
&= \mathrm{E}_{\boldsymbol{z}}\big[\log p(\boldsymbol{z}|\boldsymbol{\omega})\big] + \log p(\boldsymbol{\omega}) + \text{const} \\
&= \sum_{i=1}^{n}\sum_{q=1}^{K} \tau_{iq} \log \omega_q + \sum_{q=1}^{K} (\rho_q^0 - 1) \log \omega_q + \text{const} \\
&= \sum_{q=1}^{K} \Big(\rho_q^0 - 1 + \sum_{i=1}^{n} \tau_{iq}\Big) \log \omega_q + \text{const}
\end{aligned} \tag{15}$$

After taking the exponential of Eq. (15) and normalizing, we obtain the approximate distribution $q(\boldsymbol{\omega})$ in Eq. (14) of the parameter $\boldsymbol{\omega}$. □

Proposition 3. *Given the approximate distribution $q(\boldsymbol{z})$, the optimal approximate distribution $q(\boldsymbol{\pi})$ of the parameter $\boldsymbol{\pi}$ is the following distribution, each factor of which is a Dirichlet distribution.*

$$q(\boldsymbol{\pi}) = \prod_{l,q} \mathrm{Dir}(\boldsymbol{\pi}_{lq}; \boldsymbol{\eta}_{lq}) \tag{16}$$

For $q \neq l$, the hyperparameter η_{qlh} (h={1, 2, 3}) is given by

$$\eta_{lqh} = \eta_{lqh}^0 + \sum_{i \neq j}^{n} \tau_{il}\tau_{jq}\delta(a_{ij}, 2-h) \tag{17}$$

For $\forall q$, the hyperparameter η_{qqh} (h={1, 2, 3}) is given by

$$\eta_{qqh} = \eta_{qqh}^0 + \sum_{i<j}^{n} \tau_{iq}\tau_{jq}\delta(a_{ij}, 2-h) \tag{18}$$

Proof. According to variational Bayes, the optimal distribution $q(\boldsymbol{\omega})$ is derived as follows

$$\begin{aligned}
&\log q(\boldsymbol{\pi}) \\
&= \mathrm{E}_{\boldsymbol{z},\boldsymbol{\omega}}\big[\log p(N, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\omega})\big] + \text{const} \\
&= \mathrm{E}_{\boldsymbol{z}}\big[\log p(N|\boldsymbol{z}, \boldsymbol{\pi})\big] + \log p(\boldsymbol{\pi}) + \text{const} \\
&= \mathrm{E}_{\boldsymbol{z}}\Big[\sum_{i<j}^{n}\sum_{l,q}^{K} z_{il}z_{jq}\sum_{h=1}^{3}\delta(a_{ij}, 2-h)\log\pi_{lqh}\Big] + \sum_{l\leq q}^{K}\sum_{h=1}^{3}\eta_{lqh}^0\log\pi_{lqh} + \text{const} \\
&= \sum_{i<j}^{n}\sum_{l,q}^{K}\tau_{il}\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij}, 2-h)\log\pi_{lqh} + \sum_{l\leq q}^{K}\sum_{h=1}^{3}\eta_{lqh}^0\log\pi_{lqh} + \text{const} \\
&= \sum_{l<q}^{K}\sum_{i\neq j}^{n}\tau_{il}\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij}, 2-h)\log\pi_{lqh} \\
&\quad + \sum_{q=1}^{K}\sum_{i<j}^{n}\tau_{iq}\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij}, 2-h)\log\pi_{qqh} + \sum_{l\leq q}^{K}\sum_{h=1}^{3}\eta_{lqh}^0\log\pi_{lqh} + \text{const} \\
&= \sum_{l<q}^{K}\Big(\eta_{lqh}^0 + \sum_{i\neq j}^{n}\tau_{il}\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij}, 2-h)\Big)\log\pi_{lqh} \\
&\quad + \sum_{q=1}^{K}\Big(\eta_{qqh}^0 + \sum_{i<j}^{n}\tau_{iq}\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij}, 2-h)\Big)\log\pi_{qqh} + \text{const}
\end{aligned} \tag{19}$$

Then we take the exponential of Eq. (19) and normalize it, we can obtain the optimal distribution $q(\boldsymbol{\pi})$ in Eqs. (16)–(18) of the parameter $\boldsymbol{\pi}$. □

At this point, we have deduced all of the expressions, Eqs. (12), (14), (17) and (18), which build the main steps of our algorithm, VBS. We iterate to update these equations to convergence. Accordingly, the values of the learned hyperparameters $(\boldsymbol{\tau}, \boldsymbol{\eta})$ are used to analyze the structure of signed networks.

2.3. Time complexity analysis and a fast version

For the proposed algorithm, VBS, the time complexity is mainly determined by calculating Eqs. (12), (14), (17) and (18). Updating the τ_i of each node according to Eq. (12) takes $O(K^2 n)$, accordingly, for all the nodes, the time complexity is $O(K^2 n^2)$. According to Eq. (14), the time complexity of calculating $\boldsymbol{\rho}$ is $O(nK)$. According to Eqs. (17) and (18), updating the $\boldsymbol{\mu}$ takes $O(K^2 n^2)$. As a result, the total time complexity of the algorithm is $O(K^2 n^2)$.

To efficiently solve the large signed networks, here, we propose a fast version of the VBS. First, we rewrite the Eqs. (12), (17) and (18). For Eq. (12), it can be rewritten as follows

$$\tau_{il} \propto e^{\psi(\rho_l)-\psi(\sum_{l=1}^{K}\rho_l)} \prod_{q=1}^{K} \left(e^{\sum_{h=1}^{3} \tau_q^{(ih)} (\psi(\eta_{lqh})-\psi(\sum_{h=1}^{3}\eta_{lqh}))} \right), \tag{20}$$

where $\tau_q^{(i1)}, \tau_q^{(i2)}, \tau_q^{(i3)}$ is the sum of τ_{jq} of nodes which are connected to the node i with the positive, no and negative links, respectively.

As we can see in Eq. (20), calculating $\tau_q^{(i1)}$ takes $O(d_{i+})$, where d_{i+} is the number of positive links connected to node i. Calculating $\tau_q^{(i3)}$ takes $O(d_{i-})$, where d_{i-} is the number of negative links connected to node i. Calculating $\tau_q^{(i3)}$ takes $O(1)$ since we can calculate it by $\tau_{sum} - \tau_q^{(i1)} - \tau_q^{(i3)}$, and τ_{sum} only need be calculated once in the initialization, then updating τ_{sum} after updating τ_i. Consequently, updating τ_{il} takes $O(Kd_i)$, where $d_i = d_{i+} + d_{i-}$. For all the nodes, the time complexity is $O(K^2 d)$, where $d = \sum_{i=1}^{n} d_i$. Since $d = 2E$, where E is the number of positive and negative links in the networks, the time complexity of calculating $\boldsymbol{\tau}$ is $O(K^2 E)$.

For Eqs. (17) and (18), they can be rewritten as follows. for $q \neq l$,

$$\begin{aligned}
\eta_{lq1} &= \eta_{lq1}^0 + \sum_{i,j \in link_p}^{E_1} \tau_{il}\tau_{jq} \\
\eta_{lq3} &= \eta_{lq3}^0 + \sum_{i,j \in link_n}^{E_3} \tau_{il}\tau_{jq} \\
\eta_{lq2} &= \eta_{lq2}^0 - \eta_{lq1}^0 - \eta_{lq3}^0 - \eta_{lq1} - \eta_{lq3} + \sum_{i=1}^{n}\tau_{il} \times \sum_{i=1}^{n}\tau_{jq} - \sum_{i=1}^{n}\tau_{jl}\tau_{jq}
\end{aligned} \tag{21}$$

for $\forall q$,

$$\begin{aligned}
\eta_{qq1} &= \eta_{qq1}^0 + \frac{1}{2}\sum_{i \in link_p}^{E_1} \tau_{iq}^2 \\
\eta_{qq3} &= \eta_{qq3}^0 + \frac{1}{2}\sum_{i \in link_n}^{E_3} \tau_{iq}^2 \\
\eta_{qq2} &= \eta_{qq2}^0 - \eta_{qq1}^0 - \eta_{qq3}^0 - \eta_{qq1} - \eta_{qq3} + \frac{1}{2}\left(\sum_{i=1}^{n}\tau_{il} \times \sum_{i=1}^{n}\tau_{jq} - \sum_{i=1}^{n}\tau_{jq}\tau_{jq} \right),
\end{aligned} \tag{22}$$

where $link_p$ and $link_n$ denote the set of positive and negative links in the signed network, respectively. For each undirected link, there are two links in the link sets, for example, (i, j) and (j, i). E_1 and E_3 respectively denote the number of links in $link_p$ and $link_n$, and $E_1 + E_3 = 2E$. Accordingly, calculating η_{lq} in Eq. (21) takes $O(E)$, calculating η_{qq} in Eq. (22) takes $O(E)$. So, calculating $\boldsymbol{\eta}$ takes $O(K^2 E)$.

For the fast version, we need to iterate to update Eqs. (14), (20)–(22). Consequently, the time complexity of the fast algorithm is $O(K^2 E)$. For large sparse networks, usually $E \ll n^2$, so that the fast algorithm can efficiently reduce the time complexity. In the experiments, the runtime is 36 s on a conventional personal computer with a 2.6 GHz CPU and 4 GB RAM for the network with 10^4 nodes, 99674 edges and 4 communities.

2.4. Evidence approximation and model selection

When the number of groups is given, we can directly run the above algorithm in the signed networks. However, when the number of groups is unknown, we also need to find the optimal model. This is the problem of model selection. In the Bayesian framework, the evidence (marginal likelihood) can be used to select the optimal model from the model set [23].

However, computing the exact evidence could be quite difficult since we have to integrate over all possible parameter values, so that one usually find an approximation of evidence as model selection criterion, such as BIC, one popular approximation. In the variational Bayesian framework, the variational lower bound of the evidence can be used for Bayesian model selection [22,23], but this requires to deduce the expression of variational lower bound for the specific model.

Next, we deduce the expression of the variational lower bound of the proposed model. We also give a specific model selection criterion for the proposed model based on its variational lower round.

Proposition 4. *The lower bound of the evidence of the proposed model only depends on the posterior probability* $\boldsymbol{\tau}$ *and the normalizing constant of the Dirichlet distribution, and their form is as follows*

$$\begin{aligned}\mathcal{L}(q(\cdot)) &= \log\left\{\frac{\Gamma(\sum_{l=1}^{K}\rho_l^0)\prod_{l=1}^{K}\Gamma(\rho_l)}{\Gamma(\sum_{l=1}^{K}\rho_l)\prod_{l=1}^{K}\Gamma(\rho_l^0)}\right\}\\ &\quad+\sum_{l\le q}\log\left\{\frac{\Gamma(\sum_{h=1}^{3}\eta_{lqh}^0)\prod_{h=1}^{3}\Gamma(\eta_{lqh})}{\Gamma(\sum_{h=1}^{3}\eta_{lqh})\prod_{h=1}^{3}\Gamma(\eta_{lqh}^0)}\right\}-\sum_{i=1}^{n}\sum_{q=1}^{K}\tau_{iq}\log\tau_{iq},\end{aligned} \tag{23}$$

where $\Gamma(\cdot)$ *is the gamma function.*

Proof. The lower bound is derived as follows

$$\begin{aligned}&\mathcal{L}(q(\cdot))\\ &=\sum_{\boldsymbol{z}}\int\int q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})\log\left\{\frac{p(N,\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})}{q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})}\right\}d_{\boldsymbol{\pi}}d_{\boldsymbol{\omega}}\\ &=\mathrm{E}_{\boldsymbol{z},\boldsymbol{\pi}}\big[\log p(N|\boldsymbol{z},\boldsymbol{\pi})\big]+\mathrm{E}_{\boldsymbol{z},\boldsymbol{\omega}}\big[\log p(\boldsymbol{z}|\boldsymbol{\omega})\big]+\mathrm{E}_{\boldsymbol{\pi}}\big[\log p(\boldsymbol{\pi})\big]+\mathrm{E}_{\boldsymbol{\omega}}\big[\log p(\boldsymbol{\omega})\big]\\ &\quad-\sum_{i}^{n}\mathrm{E}_{z_i}\big[\log q(z_i)\big]-\mathrm{E}_{\boldsymbol{\pi}}\big[\log q(\boldsymbol{\pi})\big]-\mathrm{E}_{\boldsymbol{\omega}}\big[\log q(\boldsymbol{\omega})\big]\\ &=\sum_{i<j}^{n}\left(\sum_{l,q}^{K}\tau_{il}\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij},2-h)\left(\psi\left(\eta_{lqh}\right)-\psi\left(\sum_{h=1}^{3}\eta_{lqh}\right)\right)\right)+\sum_{i=1}^{n}\sum_{l=1}^{K}\tau_{il}\left(\psi\left(\rho_l\right)-\psi\left(\sum_{l=1}^{K}\rho_l\right)\right)\\ &\quad+\sum_{l\le q}^{K}\left(\log\left(\Gamma\left(\sum_{h=1}^{3}\eta_{lqh}^0\right)\right)-\sum_{h=1}^{3}\log\left(\Gamma\left(\eta_{lqh}^0\right)\right)+\sum_{h=1}^{3}\left(\eta_{lqh}^0-1\right)\left(\psi\left(\eta_{lqh}\right)-\psi\left(\sum_{h}^{3}\eta_{lqh}\right)\right)\right)\\ &\quad+\log\left(\Gamma\left(\sum_{l=1}^{K}\rho_l^0\right)\right)-\sum_{l=1}^{K}\log\left(\Gamma\left(\rho_l^0\right)\right)+\sum_{l=1}^{K}\left(\rho_l^0-1\right)\left(\psi\left(\rho_l\right)-\psi\left(\sum_{l=1}^{K}\rho_l\right)\right)-\log\left(\Gamma\left(\sum_{l=1}^{K}\rho_l\right)\right)\\ &\quad+\sum_{l=1}^{K}\log\left(\Gamma\left(\rho_l\right)\right)-\sum_{l=1}^{K}\left(\rho_l-1\right)\left(\psi\left(\rho_l\right)-\psi\left(\sum_{l=1}^{K}\rho_l\right)\right)-\sum_{l\le q}^{K}\left(\log\left(\Gamma\left(\sum_{h=1}^{3}\eta_{lqh}\right)\right)\right.\\ &\quad\left.-\sum_{h=1}^{3}\log\left(\Gamma\left(\eta_{lqh}\right)\right)+\sum_{h}^{3}\left(\eta_{lqh}-1\right)\left(\psi\left(\eta_{lqh}\right)-\psi\left(\sum_{h=1}^{3}\eta_{lqh}\right)\right)\right)-\sum_{i=1}^{n}\sum_{l=1}^{K}\tau_{il}\log(\tau_{il})\\ &=\sum_{l\le q}^{K}\sum_{h=1}^{3}\left(\eta_{lqh}^0-\eta_{lqh}+\sum_{i<j}^{n}\tau_{iq}\tau_{jq}\delta(a_{ij},2-h)\right)\times\left(\psi\left(\eta_{lqh}\right)-\psi\left(\sum_{h=1}^{3}\eta_{lqh}\right)\right)\\ &\quad+\sum_{l=1}^{K}\left(\left(\rho_l^0-\rho_l+\sum_{i=1}^{n}\tau_{il}\right)\left(\psi\left(\rho_l\right)-\psi\left(\sum_{l=1}^{K}\rho_l\right)\right)\right)+\log\left\{\frac{\Gamma\left(\sum_{l=1}^{K}\rho_l^0\right)\prod_{l=1}^{K}\Gamma\left(\rho_l\right)}{\Gamma\left(\sum_{l=1}^{K}\rho_l\right)\prod_{l=1}^{K}\Gamma\left(\rho_l^0\right)}\right\}\\ &\quad+\sum_{l\le q}^{K}\log\left\{\frac{\Gamma(\sum_{h=1}^{3}\eta_{lqh}^0)\prod_{h=1}^{3}\Gamma(\eta_{lqh})}{\Gamma(\sum_{h=1}^{3}\eta_{lqh})\prod_{h=1}^{3}\Gamma(\eta_{lqh}^0)}\right\}-\sum_{i=1}^{n}\sum_{l=1}^{K}\tau_{il}\log\tau_{il}\end{aligned} \tag{24}$$

According to Eqs. (12), (14), (17) and (18), the terms $\eta_{lqh}^0-\eta_{lqh}+\sum_{i<j}^{n}\tau_{iq}\tau_{iq}\delta(a_{ij},h)$, and $\rho_q^0-\rho_q+\sum_{i=1}^{n}\tau_{iq}$ in the lower bound vanish. Finally the lower bound of the evidence in Eq. (23) is obtained. □

However, for the proposed model, when the lower bound is used for model selection, it needs to be modified somewhat to take into account the lack of identifiability of the parameters. Although the variational Bayes will approximate the volume occupied by the parameter posterior, it will only do so around one of the local modes. With K groups, there are $K!$ equivalent modes, which differ merely by permuting the labels. Therefore we should use $L(q(\cdot))+\log(K!)$ for model comparison.

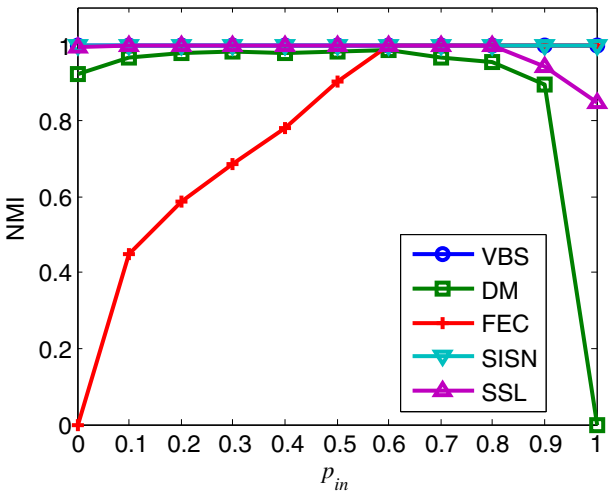

Fig. 3. Results of five algorithms in Networks I.

Consequently, the proposed model selection criterion for our model, VBS_c, can be written as follows

$$VBS_c = \log\left\{\frac{\Gamma(\sum_{l=1}^{K}\rho_l^0)\prod_{l=1}^{K}\Gamma(\rho_l)}{\Gamma(\sum_{l=1}^{K}\rho_l)\prod_{l=1}^{K}\Gamma(\rho_l^0)}\right\} + \sum_{l\le q}^{K}\log\left\{\frac{\Gamma(\sum_{h=1}^{3}\eta_{lqh}^0)\prod_{h=1}^{3}\Gamma(\eta_{lqh})}{\Gamma(\sum_{h=1}^{3}\eta_{lqh})\prod_{h=1}^{3}\Gamma(\eta_{lqh}^0)}\right\} - \sum_{i=1}^{n}\sum_{q=1}^{K}\tau_{iq}\log\tau_{iq} + \log(K!) \tag{25}$$

In addition, there are other two advantages to calculate the lower bound of our model. First, it can be used to assess convergence of the VBS. Second, it can be used to validate the correctness of the code of the VBS, if the lower bound does not increase monotonically, there will be the errors in the code of the VBS.

3. Experiments

The proposed algorithm, VBS, is validated in the synthetic and real-world networks in our experiments. We also make comparisons with other four algorithms which are respectively DM [13], SSL [1], FEC [16] and SISN [2]. And use the normalized mutual information (NMI) [2] to evaluate the performance of the algorithms. The range for the NMI value is from 0 to 1. The larger the NMI value is, the better the performance of the algorithm is.

3.1. Synthetic signed networks

In our experiments, most of the synthetic networks are generated by the generate model in Ref. [16], which is defined as follows

$$Model = M(c, n, k, p_{in}, p-, p+), \tag{26}$$

where c, n and k respectively denote the number of communities, the number of nodes in each community and the average degree of the nodes, p_{in} is the probability of the node connecting to other nodes in the same community, accordingly, $1 - p_{in}$ is the probability of the node connecting to other nodes in the different communities, $p-$ and $p+$ are respectively the probability of negative links within communities and positive links between communities, which are also called the noise parameters.

To efficiently validate our algorithm, we generate six types of signed networks with different structures in our experiments. Among them, only one type of networks (Networks I) is balanced, the others are unbalanced.

Networks I: Generated by the model with specific parameters (4, 32, 32, p_{in}, 0, 0) and varying p_{in} from 0 to 1 with the interval 0.1. The type of signed networks is balanced, in which all the positive links lie in the communities and all the negative links are between the communities. They are also called the networks without noises in the literature [16].

We run the five algorithms in these networks and the results are shown in Fig. 3. As we see in Fig. 3, the VBS and SISN can correctly find all the communities. This indicates the VBS and SISN are insensitive to the change of the parameter p_{in}, and both of them have the excellent performance in the balanced networks. The SSL also shows the good performance, it can correctly find the communities when $p_{in} <= 0.8$. **Networks II:** Generated by the model with specific parameters $(4, 32, 32, 0.5, 0.05 * p-, 0)$ and varying $p-$ from 0 to 0.5 with the interval 0.05. The type of networks is unbalanced, but

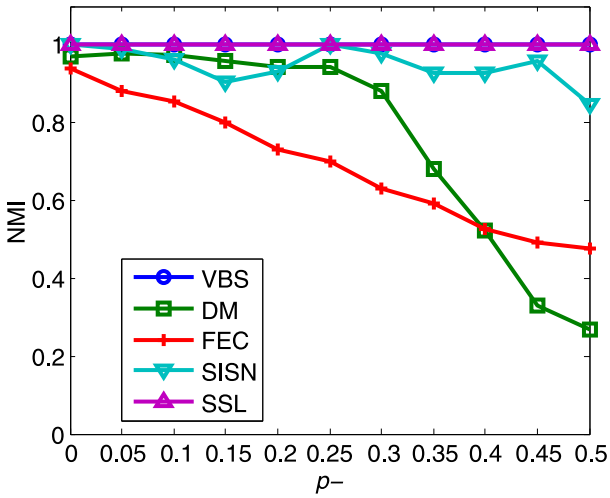

Fig. 4. Results of five algorithms in Networks II.

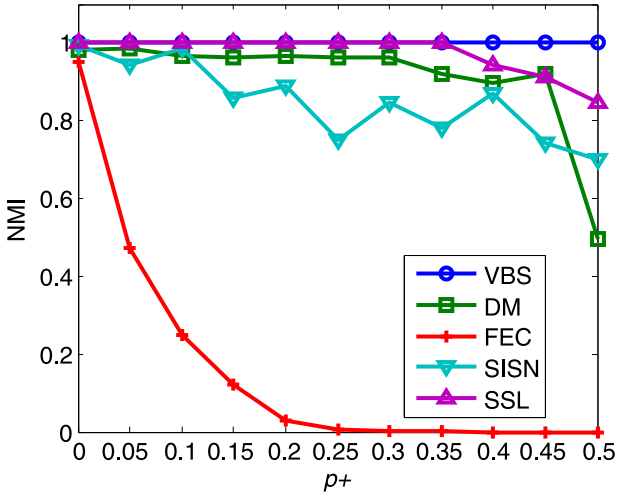

Fig. 5. Results of five algorithms in Networks III.

the noises only lie in the communities. There are not the positive links between the communities. The larger the value of $p-$ is, the more the negative links in the communities are.

The results of five algorithms are shown in Fig. 4. We can see that, all the NMI values of the VBS and SSL are 1 when $p\_$ varies from 0 to 0.5. This indicates our method and SSL can correctly find the communities in the networks. **Network III:** Generated by the model with specific parameters $(4, 32, 32, 0.5, 0, 0.05 * p+)$ and varying $p+$ from 0 to 0.5 with the interval 0.05. The type of networks is unbalanced, but the noises only lie between the communities, there are not the negative links in the communities. The larger the value of $p+$ is, the more the positive links between the communities are.

The results of five algorithms are shown in Fig. 5. As we can see in Fig. 5, the VBS can correctly find the communities, but the performance of the SSL begins to decline when $p+ > 0.35$. The VBS shows the best performance among five algorithms. **Network IV:** Generated by the model with specific parameters $(4, 32, 32, 0.5, 0.05 * p-, 0.5)$ and varying $p-$ from 0 to 0.5 with the interval 0.05. The type of networks is unbalanced, the noises lie not only in the communities but between the communities. That means there are some negative links in the communities and some positive links between the communities. The larger the value of $p-$ is, the more the negative links in the communities are.

The results of five algorithms are shown in Fig. 6. As we can see, the VBS can correctly find the communities when $p- < 0.3$, the accuracy begins to decline when $p- > 0.3$. The SSL also shows the good performance, but the DM and FEC show very bad performance. **Network V:** Generated by the model with specific parameters $(4, 32, 32, 0.5, 0.5, 0.05 * p+)$ and varying $p+$ from 0 to 0.5 with the interval 0.05. The type of networks is unbalanced, the larger the value of $p+$ is, the more the positive links between the communities are.

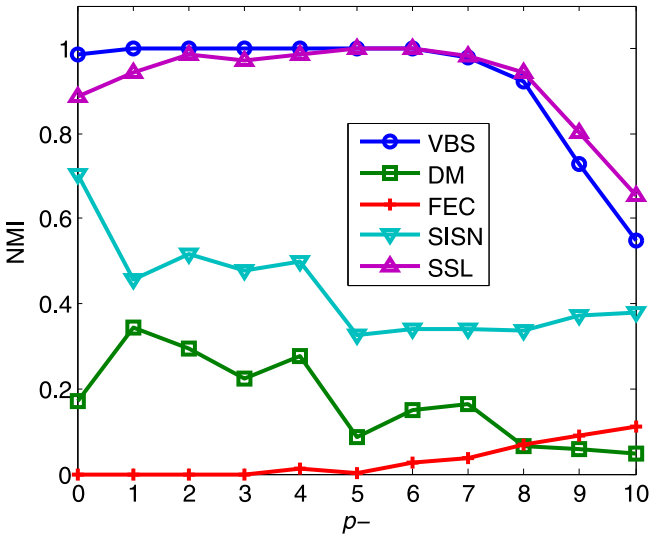

Fig. 6. Results of five algorithms in Networks IV.

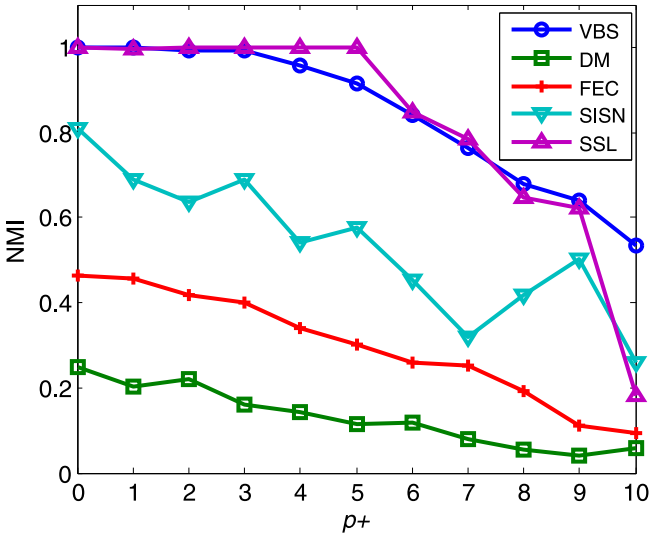

Fig. 7. Results of five algorithms in Networks V.

The results of five algorithms are shown in Fig. 7. As we can see, the VBS and SSL have the best performance among five algorithms. In contrast to other types of networks, the proposed algorithm shows the best performance in the Networks V.

Network VI: Unlike above types of networks only with communities, Network VI includes the coexisting structure with the community and bipartite structure. The type of networks is generated according to the following way: First, all the nodes are divided into four groups, each of which includes 32 nodes. Then, the links within or between the groups are generated according to the following connection probability matrix.

$$\begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} \\ & \pi_{22} & \pi_{23} & \pi_{24} \\ & & \pi_{33} & \pi_{34} \\ & & & \pi_{44} \end{bmatrix},$$

where π_{11}={0.6, 0.1, 0.3}, π_{12}={0.1, 0.2, 0.7}, π_{13}={0.1, 0.2, 0.7}, π_{14}={0.1, 0.2, 0.7}, π_{22}={0.2, 0.1, 0.7}, π_{23}={0.01, 0.4, 0.59}, π_{24}={0.01, 0.4, 0.59}, π_{33}={0.01, 0.01, 0.98}, π_{34}={0.01, 0.4, 0.59}, π_{44}={0.01, 0.01, 0.98}. The positive, no and negative links between two nodes within or between groups follow the multinomial distribution with parameter π. Fig. 8 illustrates the adjacency matrix of randomly generated network according to the above parameter set.

The results of five algorithms are shown in Figs. 9–12. The NMI values of the results for the VBS, DM, FEC, SISN and SSL are 1, 0.8641, 0, 0.8827 and 0.8338, respectively. Figs. 10–13 show the rearranged adjacency matrices of the results for five

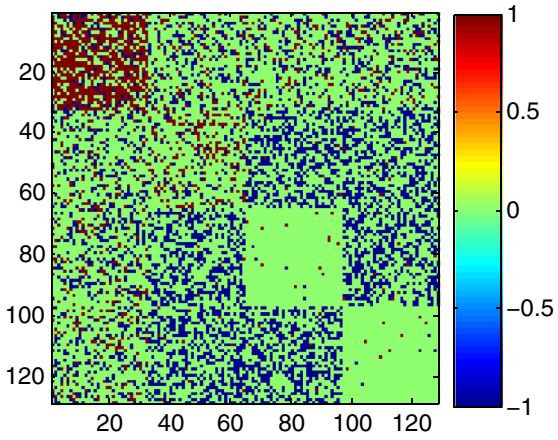

Fig. 8. Adjacency matrix of an example of Network VI.

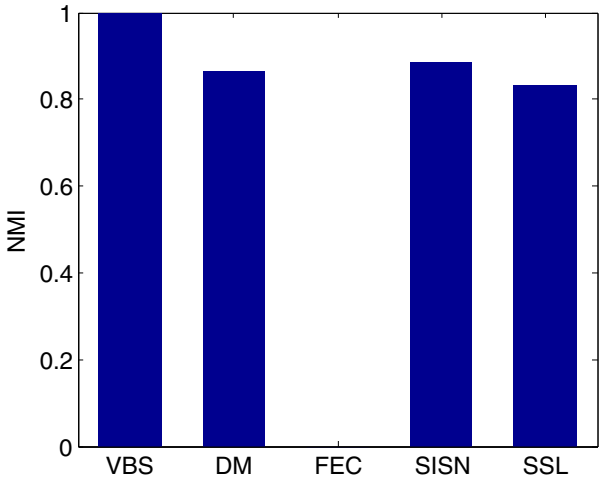

Fig. 9. Results of five algorithms in Networks VI.

Table 1
Mean of NMI value of five algorithms.

	VBS	DM	FEC	SISN	SSL
Network I	**1**	0.8746	0.7646	**1**	0.9805
Network II	**1**	0.7684	0.6933	0.9483	**1**
Network III	**1**	0.9112	0.1666	0.8511	0.9731
Network IV	**0.9245**	0.1704	0.0312	0.4315	0.9230
Network V	**0.8459**	0.1307	0.2993	0.5358	0.8259
Network VI	**1**	0.8641	0	0.8827	0.8338

algorithms, where the lines are used to separate different groups. This indicates the VBS has more excellent performance in such networks with the coexisting structure than other four algorithms.

To intuitively demonstrate the superiority of the proposed method, we list the mean of NMI values on in Table 1. The mean value is from the result of each algorithm running on the aforementioned six classes of networks. As we can see in Table 1, our proposed algorithm shows the best performance among five algorithms.

Finally, we list in detail the similarities and differences among the SISN, SSL and VBS in Table 2. In Table 2, 'EM' denotes 'Expectation Maximization', 'VBEM' denotes 'Variational Bayes EM', 'EA' denotes 'Evidence Approximation'.

3.2. Real-world signed networks

For the real-world signed networks, we select two real-world networks with ground truth community structure and three real-world networks without ground truth structure to validate the proposed algorithm. Two signed networks with ground

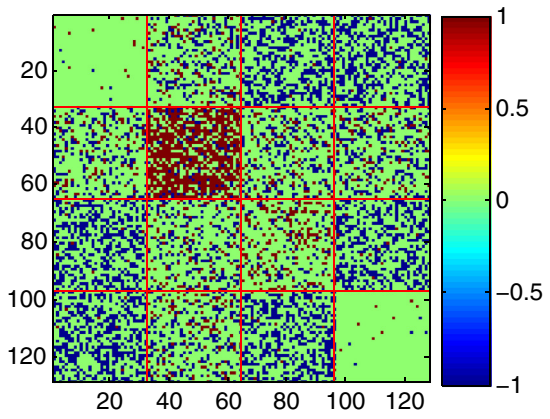

Fig. 10. Results of the VBS.

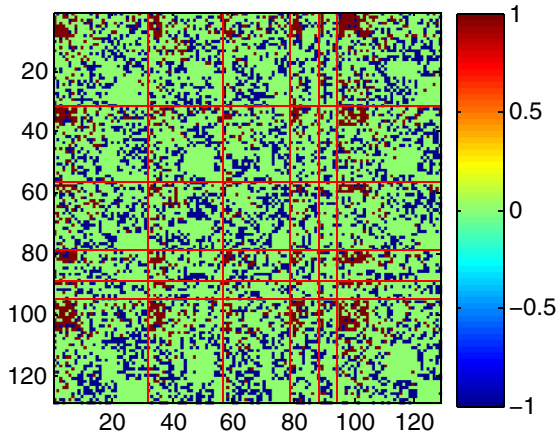

Fig. 11. Results of the DM.

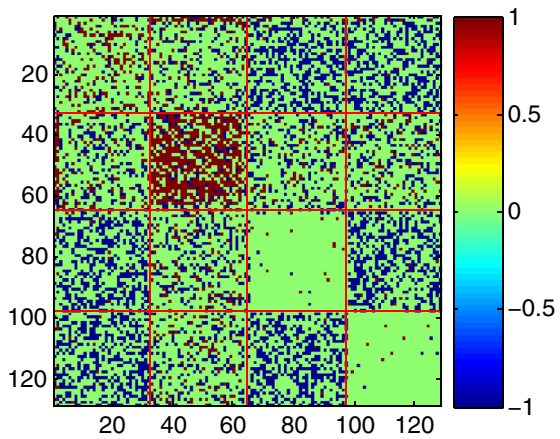

Fig. 12. Results of the SISN.

Table 2
Comparisons among the SISN, SSL and VBS.

Method	Connection pattern	Modelling structure	Number of parameters	Parameter estimation	Model selection
SISN	Block to node	Community, multipartite, mixture structure	$K(n+1)$	EM	MDL
SSL	Block to block	Community, bipartite	$K+2$	VBEM	EA
VBS	Block to block	Community, multipartite, mixture structure	$K(K+1)$	VBEM	EA

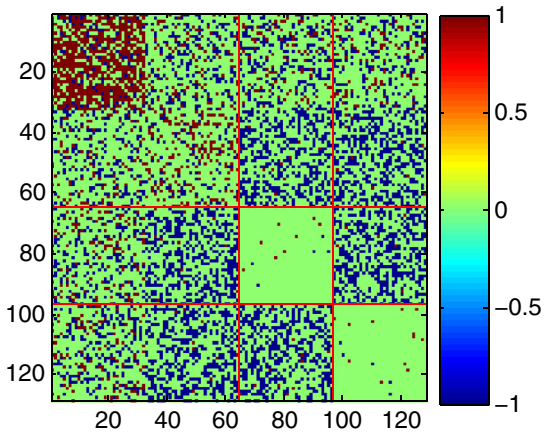

Fig. 13. Results of the SSL.

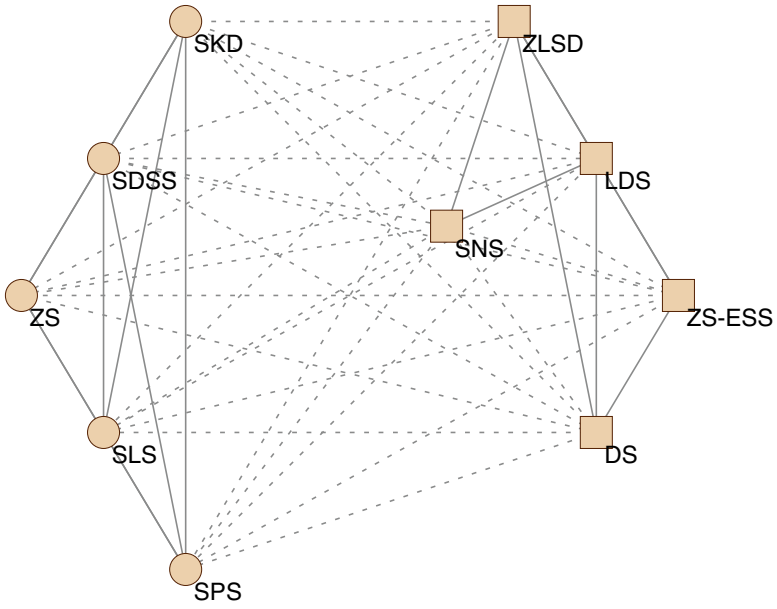

Fig. 14. Results in SPPN of the VBS.

truth community are Slovene parliamentary party network (SPPN) [25] and Gahuku-Gama subtribes network (GGSN) [26], respectively. Three signed network without ground truth structure is the Country network [27], Epinions network [10] and Slashdot network [10].

The SPPN is built according to the political relationships among 10 parties of the Slovene Parliamentary in 1994. The positive and negative links respectively describe the similar and dissimilar relationships between the parties. The GGSN describes the political relationships among 16 Gahuku-Gama subtribes in 1954, in which the positive and negative links respectively denote alliance and opposition relationships among the subtribes.

For our algorithm, we set $k_{\min} = 2$ and $k_{\max} = 8$. Fig. 14 shows the results of our algorithm in the SPPN. As we can see in Fig. 14, the nodes in the SPPN are divided into two communities. The circle nodes belong to one community and the square nodes belong to the other community. All the positive links lie in the communities and all the negative links lie between the communities. The SPPN is balanced and the result is consistent with the ground truth of SPPN. In addition, this also indicates the proposed model selection criterion is efficient.

Fig. 15 shows the results of our algorithm in the GGSN. In Fig. 15, the nodes with the same shape belong to the same community. The nodes in the GGSN are divided into three communities. We can see that, all the negative links lie between the communities and most of positive links lie in the communities. This indicates the GGSN is unbalanced. The result of our algorithm is consistent with the ground truth of the GGSN.

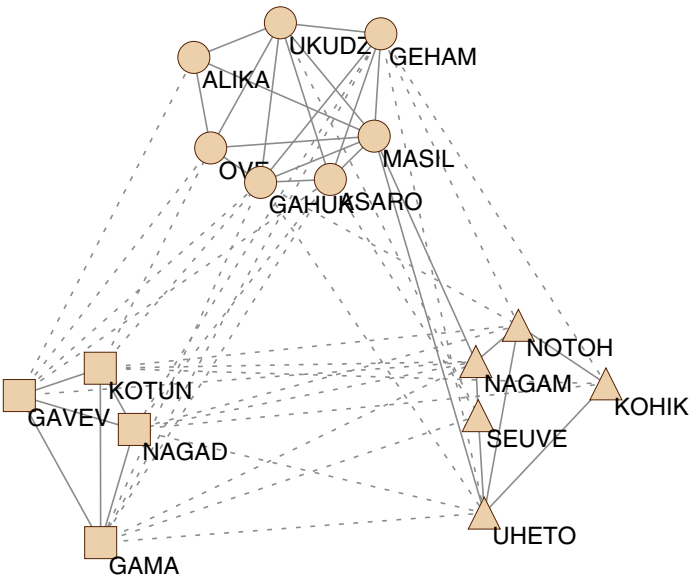

Fig. 15. Results in GGSN of the VBS.

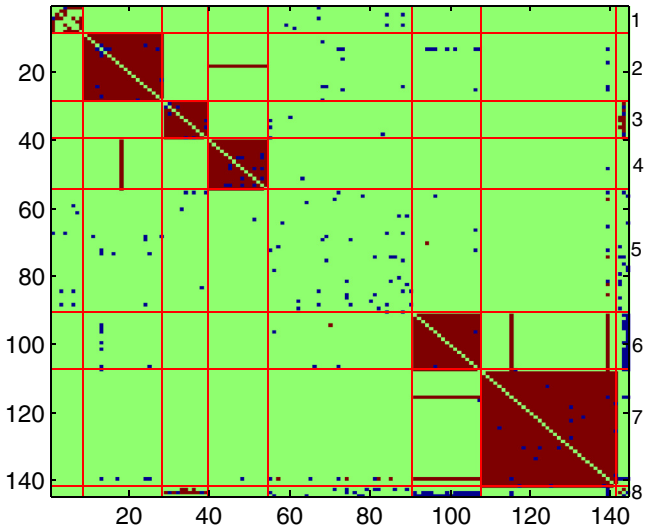

Fig. 16. Result in the country network of the VBS. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

For the Country network from the Correlates of War data set over the period 1996–1999 [27], the nodes denote the countries, the positive links represent the military alliances and the negative links denote the military disputes. In our experiments, we delete the isolated nodes and small connected component in the network, only remain the largest connected component with 144 nodes and 1243 links.

Our algorithm divides all the nodes into eight groups. The detailed results are shown in Figs. 16 and 17. Fig. 16 shows the rearranged adjacency matrix according to the group information. In Fig. 16, the red dots denote the positive links and the blue dots represent the negative links, respectively. Eight groups of nodes are separated by red solid lines. Fig. 17 intuitively shows the structure of the network. In Fig. 17, the small circles denote the nodes, the solid lines denote the positive links and the dotted lines denote the negative links. The nodes with the same color belong to the same group and form the bigger circle. The number near the big circles is the label of the groups which correspond to the number on the right in Fig. 16.

In Figs. 16 and 17, we can see that, (1) There are five communities with dense positive links in the eight groups, which are the group 2, 3, 4, 6 and 7, respectively. (2) There are two communities with the relatively sparse positive links in eight groups, which are the group 1 and 8. (3) The group 5 consists of lots of peripheral nodes, in which there are few links.

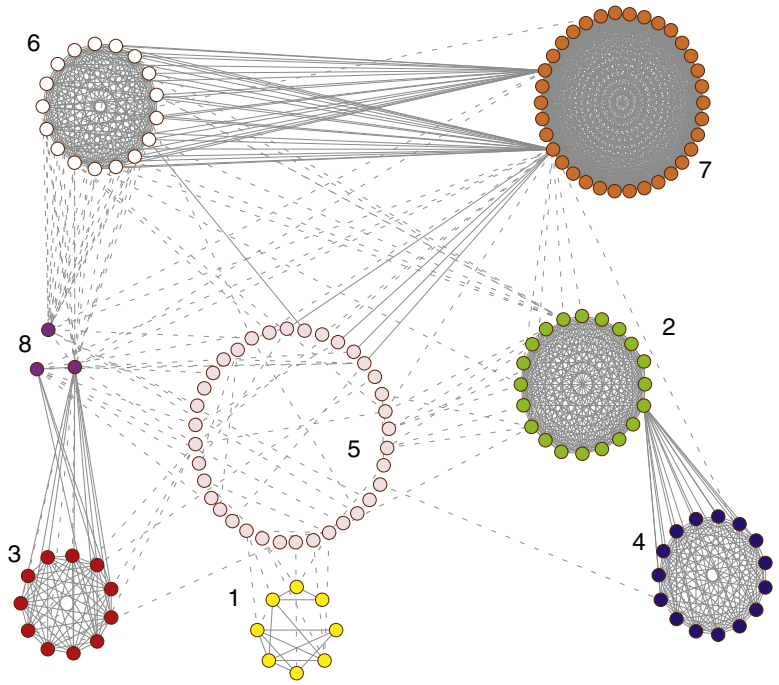

Fig. 17. Result in the country network of the VBS. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

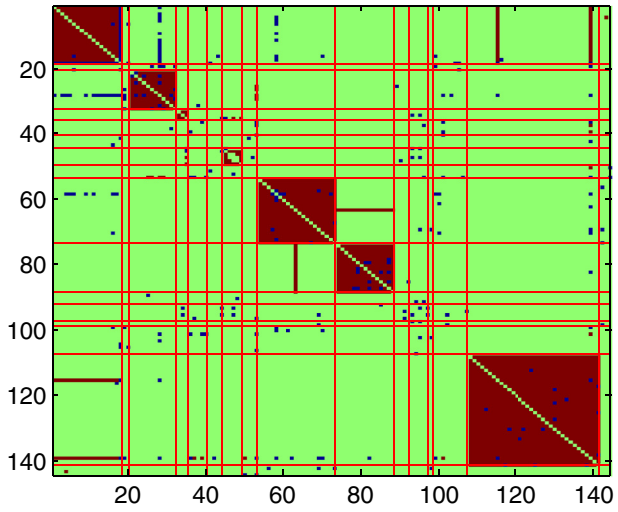

Fig. 18. Result of the SSL in the country network.

(4) The negative links mainly lie between groups. (5) For the community 2 and 4, there are no negative links between them, but there are lots of positive links between one node in the community 2 and the nodes in the community 4. (6) For community 6 and 7, there are few links between them, but there are lots of positive links between two nodes in the community 7 and the nodes in the community 6.

We also make comparisons with the SSL. Fig. 18 shows the result of the SSL in the country network. As we can see in Fig. 18, the result of our algorithm is obviously more reasonable than the result of the SSL.

The above results indicate that our algorithm can correctly find the coexisting structure of community and other structure. Unlike the current methods, they will provide us an inexact results since they merge the nodes of non-community structure into the other communities. As a result, for the real-world networks without any prior structure information, our proposed algorithm can help us to more efficiently analyze the structure of signed networks.

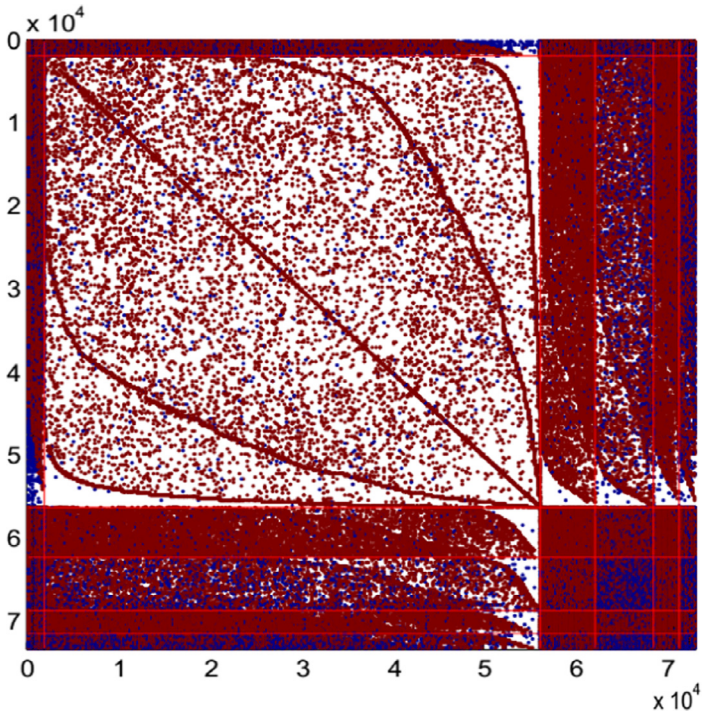

Fig. 19. Result of the VBS in the Slashdot network. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Finally, we run our algorithm in Epinions network and Slashdot network [10]. Epinions is a consumer review network in which users can either trust or distrust other consumers reviews. Slashdot is a discussion web site in which users can recognize others as friends or enemies. After the isolated nodes in the networks are deleted, the Epinions network with 126828 nodes and Slashdot network with 73099 nodes are used in our experiments. For Epinions network, sixteen groups are found by our algorithm. For Slashdot network, six groups are found by our algorithm. Fig. 19 shows the rearranged adjacency matrix of Slashdot network according to the group information. In Fig. 19, the red dots denote the positive links and the blue dots represent the negative links, respectively. Six groups of nodes are separated by red solid lines. The number of nodes in every group is 1740, 54399, 5923, 6412, 2745 and 1880, respectively. As we see in Fig. 19, the giant group with 54399 nodes is sparse, the other groups are dense. The groups with 6412 nodes and 1880 nodes contain the massive negative links, the groups with 1740 nodes, 5923 nodes and 2725 nodes contain the massive positive links.

4. Discussions and conclusions

The structure analysis is one of the important tasks in the study of the signed networks. Until now, many structure analysis methods have been proposed to analyze the signed networks, however, these methods only focus on the community structure. Recently, many researches have indicated that, most of the networks could do not consist of the single community structure, but the coexisting structure of communities and other structure. In the cases, current methods would provide us with the imprecise results of structure analysis for the signed networks. To address this problem, in this paper, we present a mathematically principled structure analysis method for the signed network. The proposed method contains two main keys that are network model and its learning algorithm. For the network model, based on block modelling idea, we propose a probability model for the signed network, which can efficiently model the well-known structure or their coexisting structure. For its learning algorithm, we deduce the specific equations of approximate posterior distributions of model parameters in the variational Bayesian framework, and a model selection criterion for the proposed model. In addition, to solve the large signed networks, we propose a fast version of the learning algorithm, which can efficiently analyze the large sparse signed networks by the general personal computers. The proposed method is validated in the synthetic and real-world signed networks and compared with current methods. The experimental results show the proposed method is more efficient to analyze the structure of the signed networks than current methods.

Finally, we discuss two main aspects of our algorithms. (1) The time complexity. Compared to the original algorithm, the time complexity of the fast version of the VBS is efficiently reduced from $O(k^2n^2)$ to $O(k^2E)$, so that the algorithm may be applied to the large sparse signed networks. It is possible to further reduce the time complexity of learning algorithm by adopting the stochastic optimization and parallel computing. This will be our future work. (2) Model selection. First, the

proposed model selection criterion, VBS_c, is efficient, but this method still requires to fit the multiple models, that means we need consider each value of k. In our experiment, we can observe that the empty groups sometimes will appear when k_{max} is set enough large, the reason is that the prior distribution of the parameter $\boldsymbol{\omega}$ adopts the Dirichlet distribution encouraging a sparse mixing weight vector. This does not affect the efficiency of the proposed model selection criterion, but it can inspire us to design a lower time complexity of learning algorithm with the ability of model selection than current methods in our future works. In addition, the code of our algorithm is available at https://github.com/xuehuazhao/network.

Acknowledgments

This work is funded by the National Science Foundation of China (Grants No. 61571444), Guangdong Province Natural Science Foundation (Grant No. 2016A030310072), Zhejiang Province Natural Science Foundation of China (Grant No. LY17F020012), Wenzhou Special Science and Technology Project (Grant No. ZG2017019), Science Research Cultivation Project of Shenzhen Institute of Information Technology (Grant No. ZY201718), Special Innovation Project of Guangdong Education Department (Grant No. 2017GKTSCX063) and MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No. 17YJCZH261).

The authors would like to thank Hechang Chen and Xu Tan for useful discussions.

References

[1] B. Yang, X. Liu, Y. Li, et al., Stochastic blockmodeling and variational Bayes learning for signed network analysis, IEEE Trans. Knowl. Data Eng. 29 (9) (2017) 2026–2039.
[2] X. Zhao, B. Yang, X. Liu, H. Chen, Statistical inference for community detection in signed networks, Phys. Rev. E 95 (4) (2017) 042313.
[3] X. Liu, W. Wang, D. He, P. Jiao, D. Jin, C.V. Cannistraci, Semi-supervised community detection based on non-negative matrix factorization with node popularity, Inf. Sci. 381 (2017) 304–321.
[4] M. Newman, Networks: an Introduction, Oxford university press, 2010.
[5] G. Ghoshal, G. Mangioni, R. Menezes, J. Poncela-Casanovas, Social system as complex networks, Soc. Netw. Anal. Min. 4 (1) (2014) 1–2.
[6] J. Chen, H. Wang, L. Wang, W. Liu, A dynamic evolutionary clustering perspective: community detection in signed networks by reconstructing neighbor sets, Phys. A Stat. Mech. Appl. 447 (2016) 482–492.
[7] S. Wang, M. Gong, H. Du, L. Ma, Q. Miao, W. Du, Optimizing dynamical changes of structural balance in signed network based on memetic algorithm, Soc. Netw. 44 (2016) 64–73.
[8] J. Tang, Y. Chang, C. Aggarwal, H. Liu, A survey of signed network mining in social media, ACM Comput. Surv. (CSUR) 49 (3) (2016) 42:1–42:37.
[9] D. Cartwright, F. Harary, Structural balance: a generalization of Heider's theory., Psychol. Rev. 63 (5) (1956) 277.
[10] J. Leskovec, D. Huttenlocher, J. Kleinberg, Signed networks in social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010, pp. 1361–1370.
[11] M.E.J. Newman, Communities, modules and large-scale structure in networks, Nat. Phys. 8 (1) (2012) 25–31.
[12] J.A. Davis, Clustering and structural balance in graphs, Hum. Relat. 20 (2) (1967) 181–187.
[13] P. Doreian, A. Mrvar, A partitioning approach to structural balance, Soc. Netw. 18 (2) (1996) 149–168.
[14] N. Bansal, A. Blum, S. Chawla, Correlation clustering, Mach. Learn. 56 (1–3) (2004) 89–113.
[15] V.A. Traag, J. Bruggeman, Community detection in networks with positive and negative links, Phys. Rev. E 80 (3) (2009) 036115.
[16] B. Yang, W. Cheung, J. Liu, Community mining from signed social networks, IEEE Trans. Knowl. Data Eng. 19 (10) (2007) 1333–1348.
[17] P. Anchuri, M. Magdon-Ismail, Communities and balance in signed networks: a spectral approach, in: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2012, pp. 235–242.
[18] C. Liu, J. Liu, Z. Jiang, A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks, IEEE Trans. Cybern. 44 (12) (2014) 2274–2287.
[19] M. Gong, Q. Cai, X. Chen, L. Ma, Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition, IEEE Trans. Evolut. Comput. 18 (1) (2014) 82–97.
[20] J.-J. Daudin, F. Picard, S. Robin, A mixture model for random graphs, Stat. Comput. 18 (2) (2008) 173–183.
[21] M.E.J. Newman, E.A. Leicht, Mixture models and exploratory analysis in networks, Proc. Nat. Acad. Sci. 104 (23) (2007) 9564–9569.
[22] P. Latouche, E. Birmele, C. Ambroise, Variational Bayesian inference and complexity control for stochastic block models, Stat. Model. 12 (1) (2012) 93–115.
[23] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.
[24] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: A review for statisticians, J. Am. Stat. Assoc 112 (518) (2017) 859–877.
[25] S. Kropivnik, A. Mrvar, An analysis of the slovene parliamentary parties network, Dev. Stat. Methodol. (1996) 209–216.
[26] K.E. Read, Cultures of the central highlands, new guinea, Southwest. J. Anthropol. 10 (1) (1954) 1–43.
[27] P. Doreian, A. Mrvar, Structural balance and signed international relations, J. Soc. Struct. 16 (2015) 1–49.