Statistical inference for community detection in signed networks

Xuehua Zhao,[1,2,*] Bo Yang,[2,3,†] Xueyan Liu,[1,2,3] and Huiling Chen[4]

[1]*School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China*

[2]*Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China*

[3]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*

[4]*College of Physics and Electronic Information, Wenzhou University, Wenzhou 325035, China*

(Received 27 November 2016; revised manuscript received 11 February 2017; published 17 April 2017)

The problem of community detection in networks has received wide attention and proves to be computationally challenging. In recent years, with the surge of signed networks with positive links and negative links, to find community structure in such signed networks has become a research focus in the area of network science. Although many methods have been proposed to address the problem, their performance seriously depends on the predefined optimization objectives or heuristics which are usually difficult to accurately describe the intrinsic structure of community. In this study, we present a statistical inference method for community detection in signed networks, in which a probabilistic model is proposed to model signed networks and the expectation-maximization–based parameter estimation method is deduced to find communities in signed networks. In addition, to efficiently analyze signed networks without any *a priori* information, a model selection criterion is also proposed to automatically determine the number of communities. In our experiments, the proposed method is tested in the synthetic and real-word signed networks and compared with current methods. The experimental results show the proposed method can more efficiently and accurately find the communities in signed networks than current methods. Notably, the proposed method is a mathematically principled method.

DOI: 10.1103/PhysRevE.95.042313

I. INTRODUCTION

A network is an abstract representation of several individuals represented as a set of nodes and of their relationships represented as a set of links between the nodes. Currently, networks have been widely used as a way to represent the patterns of connections between the components of complex systems [1,2], for example, the internet, in which the nodes denote the computers and the links denote data connections [3], food webs in which the nodes denote the species in an ecosystem and the links denote predator-prey interactions [4], and so on [5–8].

In the past decade there has been a surge of interest in both empirical studies of networks and development of mathematical and computational tools for extracting insights from networks [1,2,9]. One common approach to the study of networks is to analyze the community structure of networks [10–12]. The community is a dense subnetwork within a larger network, which may correspond to a functional unit within a complex system. For example, in metabolic networks, a community can correspond to a circuit or pathway that carries out a certain function [13,14]. In social networks, a community can correspond to a common location or workplace [15].

Currently, the problem of community detection in networks is receiving wide attention [11,12] and proves to be computationally challenging. Until now, a large number of community detection methods have been proposed to address to the problem, and their representatives include the spectral methods [16], the Kernighan-Lin algorithm [17], the clique percolation method [18], the Girvan-Newman algorithm [18], the modularity-based optimization methods [19,20], the maximum flow community algorithm [21], stochastic blockmodel-based methods [22], and others. However, most of current methods are only applied to the unsigned networks in which the links describe only whether the relationships between two nodes exist or not, and it is difficult for them to find the communities in the signed networks with positive links and negative links.

In contrast to the extensively studied unsigned networks, the signed networks contain more information by extending the single relationship to the positive and negative relationships (modeled as positive links and negative links, respectively), wherein the positive links may represent like, trust, or support membership and the negative links may represent dislike, distrust, or oppose membership. Community detection in the signed networks is to find c antagonistic communities so that most positive links lie in communities and most negative links lie between communities. In this sense, a community is consistent with a cluster defined in balance theory in social science [23,24], where a strongly (or weakly) balanced network can be divided into two (or c) clusters, so that all links within clusters are positive and all links between clusters are negative. Unfortunately, the real-world signed networks are usually unbalanced due to the frustration (negative links within clusters and positive links between clusters). For this reason it is a great challenge to design an efficient community detection method for the signed networks.

With the increase of signed networks such as social networks [25] and personality and psychopathology networks [26], many methods have been proposed to find the communities in signed networks. Doreiian and Mrvar proposed a frustration-based method according to the social balance theory (referred to as DM), in which the communities are found by minimizing the sum of the negative link quantity within communities and positive link quantity between communities [27]. Then DM was improved to partition weighted signed

*lcrlc@sina.com

†ybo@jlu.edu.cn

networks by Larusso *et al.* [23]. Very similarly, Bansal *et al.* proposed a community detection algorithm by maximizing the agreement (the number of positive intracluster links and negative intercluster links) or minimizing the disagreement (the number of negative intracluster links and positive intercluster links), referred to as AG [28]. Traag *et al.* proposed a modularity-optimization-based algorithm by improving modularity function for signed networks [29]. Based on a Markov stochastic process, Yang *et al.* proposed an random-walk based method, referred to as FEC [30]. Anchuri *et al.* proposed a generalized spectral method for signed network partition [25]. Recently, multiobjective evolutionary methods have been applied to signed network decomposition [31,32]. The aforementioned community detection methods for signed networks can be regarded as the discriminative method, which divide nodes into different communities based on either predefined optimization objectives (such as modularity) or heuristics (such as a random walk model). The performances of these methods are seriously affected by the predefined objectives or heuristics.

In recent years, statistical inference methods have been receiving great interest since they can give excellent results and are mathematically principled [33,34]. In this study, we propose a statistical inference method for community detection in signed networks (referred to as SISN), in which a probabilistic model is presented to model signed networks and an expectation-maximization (EM) -based parameter estimation method is deduced to find communities in signed networks. In addition, to efficiently analyze the signed networks without any *a priori* knowledge, a model selection criterion is given to automatically determine the number of communities in signed networks. Compared with current methods, the proposed method is principled and more expected due to its efficiency.

II. MODEL AND METHOD

The proposed method, namely, SISN, mainly includes three keys, which are the network model, parameter estimation, and model selection. For the network model, we present a new probabilistic model which can efficiently model the signed networks with community structure. For parameter estimation, based on the EM algorithm, we derive the specific equations of model parameters and posterior distribution of hidden variables. The communities in signed networks can be inferred by analyzing the learned posterior distribution. The aim of model selection is determining the number of communities in the signed networks without any *a priori* knowledge. In this study, we derive a model selection criterion based on the minimum description length principle for our model. The SISN works well for both directed and undirected signed networks, but is more simpler in the directed case, so we introduce the SISN by the directed case.

A. Model

A signed network consists of n nodes which are connected by directed links. All links fall into two categories: positive links and negative links, which represent the positive or negative relationships between two individuals, respectively. Accordingly, a signed network can be represented mathematically by an adjacency matrix A where $A_{ij} = 1$ or -1 if there is an positive or negative link from node i to node j, otherwise $A_{ij} = 0$.

Assume that the nodes belong to one of c communities; the community membership between nodes and communities can be represented by the variable z_i, which indicates the community the node i belongs to. The variable z_i is usually referred as "hidden" or "missing" data in the statistical inference since they are unknown and cannot be measured directly. For the problem of community detection, the goal is to infer them from the observed network. To infer the community memberships, a probabilistic model must be proposed to model the communities and their properties, then vary the parameters of model to find the best fit to the observed network.

In this study, the proposed probabilistic model parameterizes the probability of each possible configuration of communities assignments and links as follows.

Let θ_{ri}^{+} be the probability that a directed positive link from a particular node in community r connects to node i, θ_{ri}^{-} be the probability that a directed negative link from a particular node in community r connects to node i, and θ_{ri}^{0} be the probability without any link from a particular node in community r connects to node i. And $\theta_{ri}^{+} + \theta_{ri}^{-} + \theta_{ri}^{0} = 1$. In our model, a community is a set of nodes that have similar connection patterns to other nodes.

Let π_r be the fraction of nodes in the community r, or equivalently the probability that a randomly chosen node falls into the community r. The parameters π_r and θ_{ri} satisfy the following normalization conditions:

$$\sum_{r=1}^{c} \pi_r = 1, \qquad \sum_{s\in\{+,-,0\}} \theta_{ri}^{s} = 1, \tag{1}$$

where $s \in \{1, -1, 0\}$. The quantities in our model fall into three categories: observed data A_{ij}, hidden variable z_i, and model parameters π_r, θ_{ri}^{s}.

The proposed model is also a good generative model of a signed network with community structure, which could generate a signed network according to the following steps:

(1) Assign the nodes to the communities according to multinomial distribution with parameter π.

(2) Generate a positive link, negative link, or no link from node j to node i according to multinomial distribution with parameter θ_{ri}, wherein the node j belongs to the community r.

B. Parameters estimation

In general, the standard method fitting model to a given network is likelihood maximization, in which the probability that the network is produced by the model is maximized with respect to the model parameters. The fitting problem requires us to maximize the likelihood $P(A,z|\pi,\theta)$ with respect to π and θ, which can be done by the following equations:

$$P(A,z|\pi,\theta) = P(A|z,\pi,\theta)P(z|\pi), \tag{2}$$

$$P(A|z,\pi,\theta) = \prod_{ij} (\theta_{z_i,j}^{+})^{\delta(A_{ij},1)} (\theta_{z_i,j}^{-})^{\delta(A_{ij},-1)} (\theta_{z_i,j}^{0})^{\delta(A_{ij},0)}, \tag{3}$$

$$P(z|\pi,\theta) = \prod_{i} \pi_{z_i}. \tag{4}$$

Finally,

$$P(A,z|\pi,\theta) = \prod_i \Bigg[\pi_{z_i} \prod_j (\theta^+_{z_i,j})^{\delta(A_{ij},1)} \times (\theta^-_{z_i,j})^{\delta(A_{ij},-1)} \big(\theta^0_{z_i,j}\big)^{\delta(A_{ij},0)} \Bigg], \tag{5}$$

where $\delta(x,y)$ is Kronecker function. In fact, we usually do not work with the likelihood itself but with its logarithm:

$$\begin{aligned} \ell &= \ln P(A,z|\pi,\theta) \\ &= \sum_i \Bigg[\ln \pi_{z_i} + \sum_j \ln(\theta^+_{z_i,j})^{\delta(A_{ij},1)} (\theta^-_{z_i,j})^{\delta(A_{ij},-1)} \\ &\quad \times \big(\theta^0_{z_i,j}\big)^{\delta(A_{ij},0)} \Bigg]. \end{aligned} \tag{6}$$

However, the maximization faces the difficulty because z is unknown, and this means the value of the log likelihood is also unknown. The difficulty can be solved by making a good guess at the value of z given the network A and the model parameters π,θ. More specifically, we can calculate the posterior distribution $P(z|A,\pi,\theta)$ and, based on the posterior, calculate an expected value $\bar{\ell}$ for the log likelihood by averaging over z thus:

$$\begin{aligned} \bar{\ell} &= \sum_{z_1=1}^{c} \cdots \sum_{z_n=1}^{c} P(z|A,\pi,\theta) \sum_i \Bigg[\ln \pi_{z_i} \\ &\quad + \sum_j \ln(\theta^+_{z_i,j})^{\delta(A_{ij},1)} (\theta^-_{z_i,j})^{\delta(A_{ij},-1)} \big(\theta^0_{z_i,j}\big)^{\delta(A_{ij},0)} \Bigg] \\ &= \sum_{ir} P(z_i = r|A,\pi,\theta) \Big\{ \ln \pi_{z_r} + \sum_j [\delta(A_{ij},1) \\ &\quad \times \ln \theta^+_{r,j} + \delta(A_{ij},-1) \ln \theta^-_{r,j} + \delta(A_{ij},0) \ln \theta^0_{r,j}] \Big\} \\ &= \sum_{ir} q_{ir} \Big\{ \ln \pi_{z_r} + \sum_j [\delta(A_{ij},1) \ln \theta^+_{r,j} \\ &\quad + \delta(A_{ij},-1) \ln \theta^-_{r,j} + \delta(A_{ij},0) \ln \theta^0_{r,j}] \Big\}, \end{aligned} \tag{7}$$

where to simplify the notation we have defined $q_{ir} = P(z_i = r|A,\pi,\theta)$, which is the posterior probability that the node i belongs to the community r.

The expected log likelihood is the best estimation of ℓ, and the position of its maximum is the best estimation of model parameters. However, finding the maximum still faces a problem due to the calculation of q requiring the value of π, θ and the calculation of π, θ requiring q. The problem can be solved by an iterative way that evaluates both simultaneously. This approach is the EM, algorithm which is commonly applied to the problem of missing data.

Given π,θ, and the observed network A, the posterior probabilities q of the community memberships z of nodes can be calculated according to the following equation:

$$q_{ir} = P(z_i = r|A,\pi,\theta) = \frac{P(A,z_i = r|\pi,\theta)}{P(A|\pi,\theta)}. \tag{8}$$

The factors on the right are given by summing over the possible values of z in Eq. (5):

$$\begin{aligned} &P(A,z_i = r|\pi,\theta) \\ &= \sum_{z_1=1}^{c} \cdots \sum_{z_n=1}^{c} \delta(z_i,r) P(A,z|\pi,\theta) \\ &= \sum_{z_1=1}^{c} \cdots \sum_{z_n=1}^{c} \delta(z_i,r) \prod_k \Bigg[\pi_{z_k} \\ &\quad \times \prod_j (\theta^+_{z_k,j})^{\delta(A_{kj},1)} (\theta^-_{z_k,j})^{\delta(A_{kj},-1)} \big(\theta^0_{z_k,j}\big)^{\delta(A_{kj},0)} \Bigg] \\ &= \Bigg[\pi_r \prod_j (\theta^+_{rj})^{\delta(A_{ij},1)} (\theta^-_{rj})^{\delta(A_{ij},-1)} \big(\theta^0_{rj}\big)^{\delta(A_{ij},0)} \Bigg] \\ &\quad \times \Bigg[\prod_{k\neq i} \sum_{h=1}^{c} \pi_h \prod_j (\theta^+_{hj})^{\delta(A_{kj},1)} (\theta^-_{hj})^{\delta(A_{kj},-1)} (\theta^0_{hj})^{\delta(A_{kj},0)} \Bigg] \end{aligned} \tag{9}$$

and

$$\begin{aligned} &P(A|\pi,\theta) \\ &= \sum_{z_1=1}^{c} \cdots \sum_{z_n=1}^{c} P(A,z|\pi,\theta) \\ &= \prod_k \sum_{h=1}^{c} \pi_h \prod_j \big[(\theta^+_{hj})^{\delta(A_{kj},1)} (\theta^-_{hj})^{\delta(A_{kj},-1)} \big(\theta^0_{hj}\big)^{\delta(A_{kj},0)}\big], \end{aligned} \tag{10}$$

where $\delta(x,y)$ is the Kronecker function. Substituting Eqs. (9) and (10) into Eq. (8), finally:

$$q_{ir} = \frac{\pi_r \prod_j \big[(\theta^+_{rj})^{\delta(A_{ij},1)} (\theta^-_{rj})^{\delta(A_{ij},-1)} \big(\theta^0_{rj}\big)^{\delta(A_{ij},0)}\big]}{\sum_h \pi_h \prod_j \big[(\theta^+_{hj})^{\delta(A_{ij},1)} (\theta^-_{hj})^{\delta(A_{ij},-1)} \big(\theta^0_{hj}\big)^{\delta(A_{ij},0)}\big]}; \tag{11}$$

here q_{ir} satisfies the normalization condition $\sum_{r=1}^{c} q_{ir} = 1$.

Once the value of the q is calculated, the expected log likelihood Eq. (7) can be evaluated. Accordingly, the values of π,θ can be found by maximizing Eq. (7) by introducing Lagrange multipliers to enforce the normalization conditions Eq. (1) and differentiating. Finally,

$$\pi_r = \frac{1}{n} \sum_i q_{ir}, \tag{12}$$

$$\theta^+_{rj} = \frac{\sum_i \delta(A_{ij},1) q_{ir}}{\sum_i q_{ir}}, \tag{13}$$

$$\theta^-_{rj} = \frac{\sum_i \delta(A_{ij},-1) q_{ir}}{\sum_i q_{ir}}, \tag{14}$$

$$\theta^0_{rj} = \frac{\sum_i \delta(A_{ij},0) q_{ir}}{\sum_i q_{ir}}. \tag{15}$$

Equations (11)–(15) are the keys of our algorithm. Iterating these equations to convergence, the outputs are q_{ir}, θ^+_{ri}, θ^-_{ri}, and θ^0_{ri}.

C. Model selection

At present, the specific number of communities, c, which is also regarded as *a priori* knowledge, need to be given before the above algorithm runs. However, for the real-world networks, the prior knowledge is usually unknown for us. Since there is the linear relationship between the number of communities, c, and the number of model parameters, $c(3n+1)$, how to determine the value of c is naturally the problem of model selection. Here we derive a model selection criterion based on a minimum description length principle for our model.

According to the minimum description length principle, the code length describing the network data is composed of two parts, where the first part describes the coding length of the network generated by the model, and the second part gives the length for coding all parameters of the model. The length of the coding network is the negative log likelihood, $-\ell$. To code the parameters, a precision ϵ has to be prespecified. The parameters, which are smaller than ϵ, are not coded and get a description length of zero; otherwise coding the parameters $\pi_r, \theta_{ri}^+, \theta_{ri}^-$, and θ_{ri}^0 needs $\ln(\frac{\pi_r}{\epsilon})$, $\ln(\frac{\theta_{ri}^+}{\epsilon})$, $\ln(\frac{\theta_{ri}^-}{\epsilon})$, and $\ln(\frac{\theta_{ri}^0}{\epsilon})$, respectively. As a result, the total length H for coding the model is as follows:

$$\begin{aligned} H = &-\ell + \sum_{r=1}^{c} \ln\left(\frac{\pi_r}{\epsilon}\right)\delta(\pi_r \geqslant \epsilon) \\ &+ \sum_{r=1}^{c}\sum_{i=1}^{n}\left[\ln\left(\frac{\theta_{ri}^+}{\epsilon}\right)\delta(\theta_{ri}^+ \geqslant \epsilon) + \ln\left(\frac{\theta_{ri}^-}{\epsilon}\right)\delta(\theta_{ri}^- \geqslant \epsilon)\right. \\ &\left. + \ln\left(\frac{\theta_{ri}^0}{\epsilon}\right)\delta(\theta_{ri}^0 \geqslant \epsilon)\right]. \end{aligned} \tag{16}$$

Choosing with precision ϵ is tricky but very important. Smaller ϵ may cause longer code for parameters, and hence it will always prefer models with small c. In fact, it is shown that the networks are organized in a hierarchical way, and the choice of ϵ gives a lever for viewing networks in different resolutions. The precision is usually empirical, and the results of model selection are robust to the choice of ϵ ranging from $1/0.4n$ to $1/3n$ according to our experiments. In this study, ϵ is set to $1/0.8n$, and the specific model selection criterion is the following equation:

$$\begin{aligned} H = &-\ell + \sum_{r=1}^{c} \ln\left(\frac{\pi_r}{1/0.8n}\right)\delta(\pi_r \geqslant 1/0.8n) \\ &+ \sum_{r=1}^{c}\sum_{i=1}^{n} \ln\left(\frac{\theta_{ri}^+}{1/0.8n}\right)\delta(\theta_{ri}^+ \geqslant 1/0.8n) \\ &+ \sum_{r=1}^{c}\sum_{i=1}^{n} \ln\left(\frac{\theta_{ri}^-}{1/0.8n}\right)\delta(\theta_{ri}^- \geqslant 1/0.8n) \\ &+ \sum_{r=1}^{c}\sum_{i=1}^{n} \ln\left(\frac{\theta_{ri}^0}{1/0.8n}\right)\delta(\theta_{ri}^0 \geqslant 1/0.8n). \end{aligned} \tag{17}$$

D. Time complexity analysis

For SISN, Eqs. (11)–(15) are the most time-consuming parts, as they dominate the entire learning process. According to Eq. (11), it takes $O(cn^2)$ time to calculate q where c is the number of communities and n is the number of nodes in a network. According to Eq. (12), it takes $O(cn)$ time to calculate π. According to Eqs. (13), (14), and (15), it takes the same time, $O(cn^2)$, to calculate θ^+, θ^-, and θ^0. When running the SISN, Eqs. (11)–(15) need to be iteratively calculated until convergence. Assuming the iterative times are T, when the specific number of communities, c, is given, the time complexity of the SISN is $O(Tcn^2)$. When model selection is used in the algorithm, the time calculating Eq. (17), $O(n^2)$, should also be considered; the time complexity of the SISN is $O(Tn^4)$ in the worst case.

TABLE I. Time complexity of five methods.

Algorithms	SISN	FEC [30]	DM [27]	AG [28]	SSL [35]
Complexity	$O(n^4)$	$O(n^3)$	$O(n^5)$	$O(n^5)$	$O(n^4)$

Here we make comparisons between SISN and four other methods. The time complexity of five methods is listed in Table I, where the time complexity is the time complexity of each method in the worst case. As we can see, the time complexity of the proposed method, $O(n^4)$, is the same as that of SSL. The time complexity of FEC, $O(n^3)$, is the lowest among the five methods. The time complexity of DM and AG, $O(n^5)$, is the highest among five methods.

For the SISN, if the community structure of networks is clear, the SISN would quickly converge. Here a simple example is given to intuitively explain the consumed time of the SISN. We generate a group of networks with 500 nodes and four communities by the network model (18) with the parameters $SG(4,125,300,0.5,0,0)$; the experiments are performed on a conventional personal computer with a Intel Core i5-3230M CPU and 4 GB of RAM. When $c=4$ is fixed, the average run time of the SISN in each network is 17 s. When $c_{min}=2$ and $c_{max}=10$, the average time is 312 s.

III. EXPERIMENTS

In this section, we validate the proposed method on the synthetic signed networks and real-world signed networks, and make comparisons with the other four methods: DM [27], FEC [30], AG [28], and SSL [35].

A. Synthetic networks

In our experiments, we use the model in Ref. [30] to generate synthetic networks, which is defined as follows:

$$\text{Model}_{\text{sign}} = SM(c,n,k,p_{in},p_-,p_+), \tag{18}$$

where c denotes the number of communities in the network, n is the number of nodes in each community, k is the degree of node, p_{in} is the probability of the node connecting to other nodes in the same community, accordingly, $1-p_{in}$ denotes the node connecting to other nodes in the different communities, p_- is the probability of negative links within communities, and p_+ is the probability of positive links between communities. In general, the sign of the links within communities are positive, and the sign of the links between communities are negative. For the model, p_- and p_+ also are called the noise parameters; when $p_-=0$ and $p_+=0$, that is

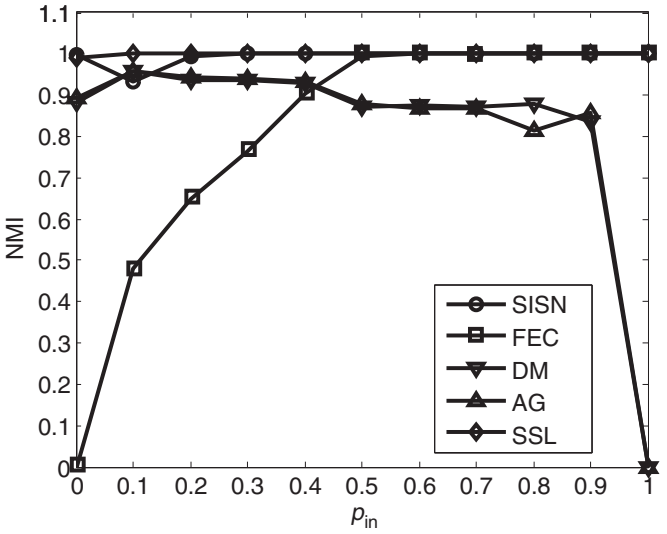

FIG. 1. Results of the five methods in synthetic signed networks without noises.

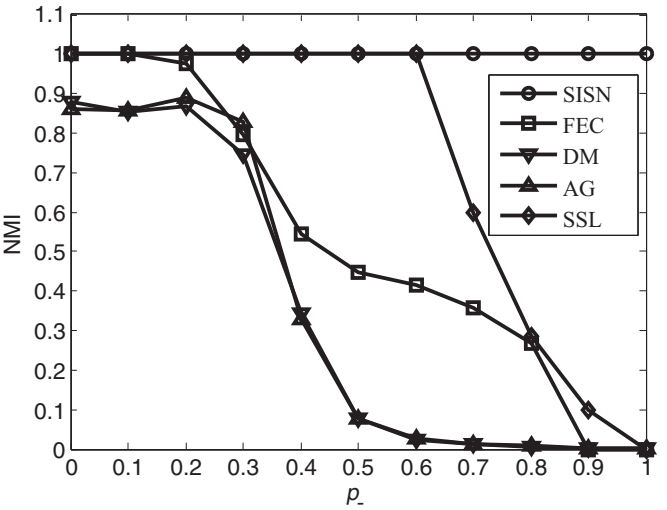

FIG. 2. Results of the five methods in the first class of synthetic signed networks with noises.

say, the generation model is $SG(c,n,k,p_{in},0,0)$, the networks are partitionable or balance. The parameter p_{in} controls the link density within the communities. When $p_{in} \to 0$, their community structures become increasingly ambiguous so that it is difficult to identify them. Given a fixed p_{in}, we can control the noise level, that is, the number of negative links between communities and the number of positive links between communities.

To evaluate the performance of the methods, the normalized mutual information (NMI) [36] is used as the evaluating measure in our experiments. Given A and B, which are the real communities and found communities, respectively, M is a confusion matrix, and m_{ij} is the number of nodes shared by the community i in A and by the community j in B. Then the NMI is defined as

$$\mathrm{NMI}(A,B) = \frac{-2\sum_{i=1}^{N_A}\sum_{j=1}^{N_B} m_{ij}\log\left(\frac{m_{ij}n}{m_{i.}m_{.j}}\right)}{\sum_{i=1}^{N_A} m_{i.}\log\frac{m_{i.}}{n} + \sum_{j=1}^{N_B} m_{.j}\log\frac{m_{.j}}{M}}, \qquad (19)$$

where $m_{i.}$ ($m_{.j}$) is the sum of elements in row i (column j), N_A and N_B are the number of communities in A and B, respectively, and n is the number of nodes in the network. NMI $= 1$ means that the found communities and real communities are identical and NMI $= 0$ indicates that the found communities and real communities are completely different. Accordingly, the better the performance of the method is, the larger the NMI is.

At first, we test the algorithms in the synthetic signed networks without noises. The signed networks are generated by the model with specific parameters $SG(4,32,32,p_{in},0,0)$ and vary p_{in} from 0 to 1 with the interval 0.1. Finally, 11 groups of networks are generated, and each group includes the 30 synthetic networks. Then we run the five methods in these networks, and the mean of NMI values are calculated. The results are shown in Fig. 1.

As we see in Fig. 1, the SISN and SSL can correctly find all the communities when $p_{in} >= 0.2$ and are insensitive to the change of parameter p_{in}. For FEC, when $p_{in} \geqslant 0.5$, it can correctly find the communities. DM and AG have nearly the same performance because their optimized objecting functions are naturally same and they cannot entirely correctly find the community. This indicates that our proposed method is more efficient than FEC, DM, and AG in signed networks without noises.

Then, we test the algorithms in the synthetic signed networks with noises. The signed networks are generated in terms of the following two ways:

(1) $SG(4,32,32,0.8,p_-,0)$ and varying p_- from 0 to 1 with the interval 0.1.

(2) $SG(4,32,32,0.8,0,p_+)$ and varying p_+ from 0 to 1 with the interval 0.1.

For each class of signed networks, we randomly generate 11 groups of networks in which each group includes the 30 synthetic networks.

For the first class of signed networks, the results of five methods are shown in Fig. 2. We can see that all the NMI values of the SISN are 1 when p_- varies from 0 to 1. This indicates our method has the best performance among the five methods. For the SSL, when $p_- > 0.5$, the performance of the algorithm begins to drop sharply. For the FEC, its performance begins to drop sharply when $p_- > 0.2$. The DM and AG have the worst performance among the five methods.

For the second class of signed networks, the results are shown in Fig. 3. As we can see in Fig. 3, the SISN still shows excellent performance, and all the NMI values are equal to one when p_+ varies from 0 to 1. The FEC can find the total communities except $p_+ = 1$. The DM and AG have the same bad performance for noises of networks.

For further validating the performance of the proposed method, we test our method in the following model parameter space:

(1) $SG(4,32,32,p_{in},0,p_+)$ and varying p_{in}, p_+ from 0 to 1 with the interval 0.1.

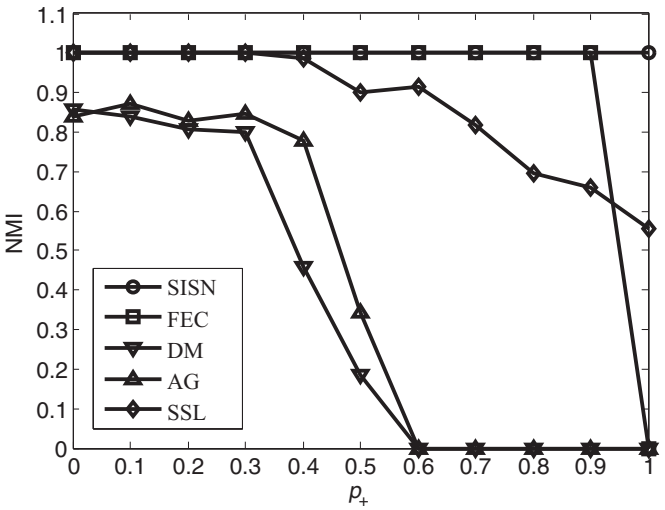

FIG. 3. Results of the five methods in the second class of synthetic signed networks with noises.

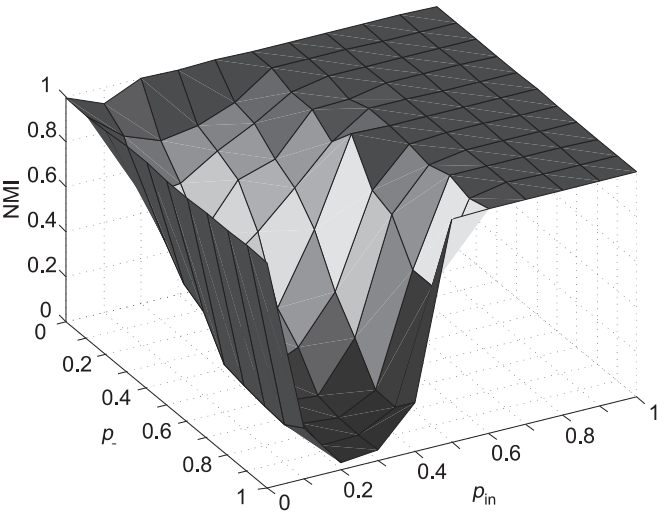

FIG. 5. Results of the SISN varying p_{in} and p_-.

(2) $SG(4,32,32,p_{in},p_-,0)$ and varying p_{in}, p_- from 0 to 1 with the interval 0.1.

Figures 4 and 5 show the performance of the SISN. As we can see in Fig. 4, when $0.1 \leqslant p_{in} \leqslant 0.4$ and $0.6 \leqslant p_+ \leqslant 1$, the performance of the proposed method is poor. Similarly, when $0.1 \leqslant p_{in} \leqslant 0.4$ and $0.6 \leqslant p_- \leqslant 1$, the performance of the proposed method is poor. This indicates that the performance of the SISN is poor only when the links within communities is very sparse and the noises are very large.

B. Real-world networks

Next, we validate the proposed method on two real-world networks: the Slovene parliamentary party network [37] and Gahuku-Gama subtribes network [38]. The Slovene parliamentary party network describes the political relationships among 10 parties in Slovenia's parliament in 1994. The positive links and negative links respectively denote that two parties are similar and dissimilar. The Gahuku-Gama subtribes network describes the political relationships among 16 Gahuku-Gama subtribes in 1954, which were distributed in highland New Guinea. The positive links and negative links in it respectively denote alliance and opposition among subtribes. The two real-world networks, which have ground truth community structures, are usually regarded as the benchmarks for testing the performance of community detection methods [27,30].

For our method, we set $c_{\min} = 2$ and $c_{\max} = 8$. Figure 6 shows the results of the SISN in the Slovene parliamentary party network. Figure 6(a) shows the network structure including two communities found by the SISN. The circle nodes belong to one community, and the square nodes belong to the other community. As we can see, all the positive links lie in the communities, and all the negative links lie between the communities. The result is consistent with the ground truth of the Slovene parliamentary party network. Figure 6(b) shows the change of the value of cost function with c. As we see, when $c = 2$, the value of the cost function is lowest and the optimal model includes two communities. This indicates the proposed model selection criterion is efficient.

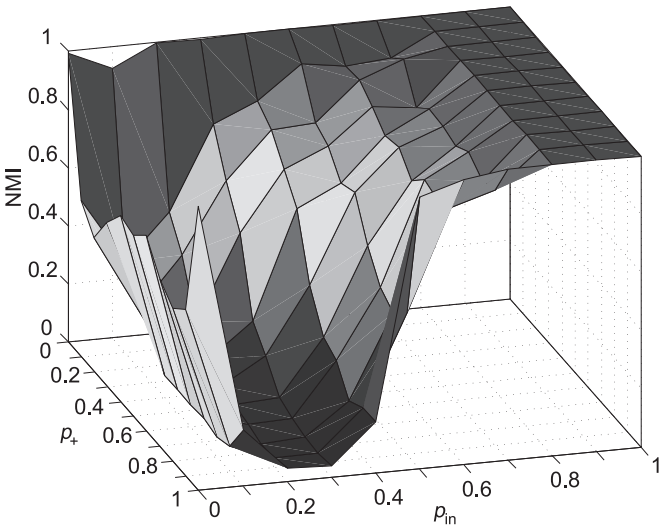

FIG. 4. Results of the SISN varying p_{in} and p_+.

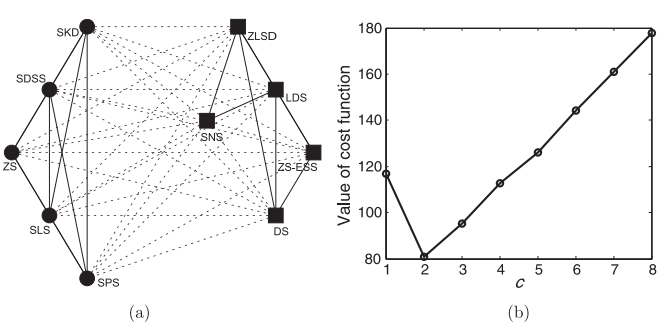

FIG. 6. Communities in the Slovene parliamentary party network found by the SISN.

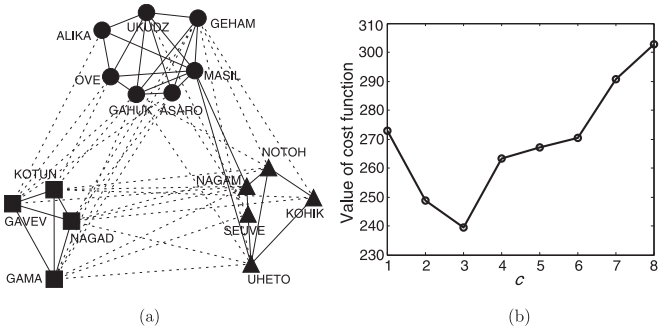

FIG. 7. Communities in the Gahuku-Gama subtribes network found by the SISN.

Figure 7 shows the results of the SISN in the Gahuku-Gama subtribes network. Figure 7(a) shows the network structure including three communities found by the SISN. The nodes with the same shape belong to the same community. As we see in Fig. 7(a), all the negative links lie between communities, and the most positive links lie in communities. The result found by the SISN is consistent with the ground truth of Gahuku-Gama subtribes network. Figure 7(b) shows the change of value of the cost function with c. We can see that when $c = 3$, the value of the cost function is the lowest and the optimal model includes three communities.

IV. CONCLUSIONS

Community detection is one of the important tasks in the analysis of signed networks. Currently, many community detection methods have been proposed to address signed networks; however, these methods can be regarded as discriminative ways because all of them depend on either predefined optimization objectives or heuristics. This is the reason that current methods have difficulty finding the intrinsic community structure.

Distinctly, in this study, we propose a probabilistic inference-based method to find the communities in signed networks, in which a probabilistic model is first presented to model the signed networks with communities, and an EM-based learning method is proposed to estimate the parameters and to infer the posterior hidden variable. In addition, a model selection criterion for the proposed model is deduced to automatically determine the number of communities in the networks. The proposed method is validated on the synthetic and real-world signed networks and compared with other methods. The experimental results show the proposed method is more efficient for finding the communities in the signed networks than other methods.

The high time complexity is the drawback of the proposed method so that it is difficult for the proposed method to efficiently analyze the large signed networks with tens of thousands of nodes. This is also an important problem of most current methods. In this study we focus more on how to propose a mathematically principled method to find the intrinsic community structure in the signed networks. We will consider adopting parallel computing and stochastic variational inference to estimate the model parameters for efficiently analyzing the large signed networks in our future work. 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. 61373053, No. 61572226, No. 61571444, and No. 61402195), Jilin Province Natural Science Foundation (Grant No. 20150101052JC), Zhejiang Province Natural Science Foundation of China (Grant No. LY17F020012), Guangdong Province Natural Science Foundation (Grant No. 2016A030310072), and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education Foundation of Jilin University (Grant No. 93K172016K19).

[1] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
[2] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, Cambridge, 2010).
[3] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, 2007).
[4] M. Pascual and J. A. Dunne, *Ecological Networks: Linking Structure to Dynamics in Food Webs* (Oxford University Press, Oxford, 2006).
[5] J. Scott, *Social Network Analysis* (Sage, Los Angeles, 2012).
[6] M. Rubinov and O. Sporns, Neuroimage **52**, 1059 (2010).
[7] D. Borsboom and A. O. Cramer, Annu. Rev. Clin. Psychol. **9**, 91 (2013).
[8] R. Guimera, S. Mossa, A. Turtschi, and L. N. Amaral, Proc. Natl. Acad. Sci. USA **102**, 7794 (2005).
[9] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, Adv. Phys. **56**, 167 (2007).
[10] M. Girvan and M. E. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[11] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[12] M. E. Newman, Nat. Phys. **8**, 25 (2012).
[13] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000).
[14] R. Guimera and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[15] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[16] Y. Bo, J. Liu, and J. Feng, IEEE Trans. Knowl. Data Eng. **24**, 326 (2012).
[17] B. W. Kernighan and S. Lin, Bell Syst. Tech. J. **49**, 291 (1970).
[18] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[19] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[20] M. E. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

[21] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, Computer **35**, 66 (2002).
[22] J.-J. Daudin, F. Picard, and S. Robin, Stat. Comput. **18**, 173 (2008).
[23] D. Cartwright and F. Harary, Psychol. Rev. **63**, 277 (1956).
[24] J. A. Davis, Human Relations **20**, 181 (1967).
[25] P. Anchuri and M. Magdon-Ismail, in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE, New York, 2012), pp. 235–242.
[26] G. Costantini and M. Perugini, PloS One **9**, e88669 (2014).
[27] P. Doreian and A. Mrvar, Social Netw. **18**, 149 (1996).
[28] N. Bansal, A. Blum, and S. Chawla, Machine Learn. **56**, 89 (2004).
[29] V. A. Traag and J. Bruggeman, Phys. Rev. E **80**, 036115 (2009).
[30] B. Yang, W. Cheung, and J. Liu, IEEE Trans. Knowl. Data Eng. **19**, 1333 (2007).
[31] C. Liu, J. Liu, and Z. Jiang, IEEE Trans. Cybernetics **44**, 2274 (2014).
[32] M. Gong, Q. Cai, X. Chen, and L. Ma, IEEE Trans. Evol. Comput. **18**, 82 (2014).
[33] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, J. Mach. Learn. Res. **9**, 1981 (2008).
[34] P. J. Bickel and A. Chen, Proc. Natl. Acad. Sci. USA **106**, 21068 (2009).
[35] B. Yang, X. Zhao, and X. Liu, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (AAAI, Menlo Park, 2015), pp. 1952–1958.
[36] L. I. Kuncheva and S. T. Hadjitodorov, in *2004 IEEE International Conference on Systems, Man and Cybernetics*, Vol. 2 (IEEE, New York, 2004), pp. 1214–1219.
[37] S. Kropivnik and A. Mrvar, Dev. Stat. Methodol., 209 (1996).
[38] K. E. Read, Southwest. J. Anthropol. **10**, 1 (1954).