Exploring the structural regularities in networks

Hua-Wei Shen,* Xue-Qi Cheng,† and Jia-Feng Guo
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
(Received 4 May 2011; revised manuscript received 24 August 2011; published 28 November 2011)

In this paper, we consider the problem of exploring structural regularities of networks by dividing the nodes of a network into groups such that the members of each group have similar patterns of connections to other groups. Specifically, we propose a general statistical model to describe network structure. In this model, a group is viewed as a hidden or unobserved quantity and it is learned by fitting the observed network data using the expectation-maximization algorithm. Compared with existing models, the most prominent strength of our model is the high flexibility. This strength enables it to possess the advantages of existing models and to overcome their shortcomings in a unified way. As a result, not only can broad types of structure be detected without prior knowledge of the type of intrinsic regularities existing in the target network, but also the type of identified structure can be directly learned from the network. Moreover, by differentiating outgoing edges from incoming edges, our model can detect several types of structural regularities beyond competing models. Tests on a number of real world and artificial networks demonstrate that our model outperforms the state-of-the-art model in shedding light on the structural regularities of networks, including the overlapping community structure, multipartite structure, and several other types of structure, which are beyond the capability of existing models.

DOI: 10.1103/PhysRevE.84.056111 PACS number(s): 89.75.Fb, 89.75.Hc, 02.50.−r, 05.10.−a

I. INTRODUCTION

Networks provide a powerful tool for representing the structure of complex systems. These networks include social networks [1,2], information networks [3,4], and biological networks [1,5]. Much of the recent research on networks actually aims to understand the structural regularities and further to reveal the relationship between such structural regularities and the function of networks [2,6]. For example, as a widely studied structural characteristic of networks, community structure is of high interest because communities often correspond to functional units, such as pathways for metabolic networks and collections of pages on a similar topic on a website.

Community structure is a kind of assortative structure, in which nodes are divided into groups such that the members within each group are mostly connected with each other. Contrary to community structure, multipartite structure is another important kind of structural regularity observed in real world networks. Multipartite structure means that nodes of the network can be divided into groups such that most of the edges are across different groups. In addition to these salient structural characteristics, other types of structure are also observed in real world networks, such as hierarchical structure and core-periphery structure.

However, existing methods mostly presume that a certain type of structure exists in the target network and are devoted to detecting such structure. This raises concerns regarding the reliability of the detected structure. On one hand, the assumed structure may not match the intrinsic structure of the target network and thus these methods are not applicable to these situations. On the other hand, several real world networks contain multiple types of structure simultaneously. Most existing methods are designed for a certain type of structure and thus cannot detect the broad types of structure. In addition, several unknown types of structure may also exist in networks and a preferred method should be able to detect such structure as well. Thus, it is important to explore multiple types of structural regularities in networks.

In the last decade, the identification of community structure has attracted much attention in various scientific fields. Many methods have been proposed and applied successfully to some specific complex networks [7–18]. For review, the reader can refer to Ref. [19]. These methods are from different perspectives, such as the centrality measures, modularity, link density, percolation theory, network compression, and spectral analysis. Recently, several generative models for network data have been proposed to detect community structure [20,21]. These models view network structure as observed quantities and take communities as hidden groups of nodes. The communities are then identified by fitting the model to the observed network structure. For example, Ren *et al.* [22] proposed a probabilistic model to uncover the overlapping community structure. This model assumes that the two end nodes of each edge are from the same community and this assumption is satisfied by the fuzzy membership of nodes. Zhang *et al.* [23] applied the latent Dirichlet allocation (LDA, a well-known generative model) to social network analysis and provided a method to detect community structure. The common drawback of these two models is that they can only uncover the community structure and fail to reveal other types of structural regularities, e.g., multipartite structure.

To characterize the hierarchical organization of networks, Clauset *et al.* proposed the hierarchical random graph model, which is capable of expressing both assortative and disassortative structure [25]. To explore more broad types of structure, Newman *et al.* proposed a mixture model for the exploratory analysis of network structure [24]. In this model, the nodes with a similar connection preference, rather than the highly connected nodes, are classified into the same group. In such a general way, this model can reveal several other kinds of

*shenhuawei@ict.ac.cn
†cxq@ict.ac.cn

structural regularities beyond community structure. However, this model fails to tell us which kind of structural regularities has been identified. More importantly, this model may produce a result that is a mixture of several types of structure, and thus the identified structure may not provide clear information about the structural regularities. The shortcoming of this model is that it only models the relationship between groups and nodes, rather than the relationship among groups. The stochastic block model provides an appropriate alternative to the mixture model for exploring a broad range of structural regularities. Karrer *et al.* utilized a degree-corrected stochastic block model [26] to investigate the community structure of a network. Airoldi *et al.* provided a mixed membership stochastic block model [27] to model network data. These works have demonstrated that the stochastic block model is a good choice for exploring the regularities of networks. However, the effectiveness of these models is limited by their inflexible model assumptions, e.g., the hard partition assumption or neglecting the directionality of edges.

In this paper, we focus on exploring the intrinsic structural regularities in networks by dividing network nodes into groups such that the members of each group have similar patterns of connections to other groups. A general stochastic block model (referred to as the GSB model in this paper) is proposed to model the network structure. In this model, node groups are represented by unobserved or hidden quantities and the relationships among groups are explicitly modeled by a block matrix as the traditional block models. Then, using the expectation-maximization algorithm, we fit the model to specific network data and detect intrinsic structural regularities of the network without prior knowledge of the type of regularity existing in the network. Compared with existing models, the most prominent strength of our model is the high flexibility. This strength enables it to possess the advantages of existing models and to overcome their shortcomings in a unified way. As a result, not only can broad types of structure be detected, but also the type of identified structure can be indicated by the block matrix. In addition, our model can tell us the centrality of the node in each group and the mixed membership of nodes as well.

Tests on a number of artificial and real world networks demonstrate that our model outperforms the state-of-the-art models in shedding light on the structural regularities of networks, including the overlapping community structure, multipartite structure, and several other types of structure, which are beyond the capability of existing models.

II. THE MODEL

Generally, a network with n nodes can be represented mathematically by an adjacency matrix A with elements $A_{ij} = 1$ if there is an edge from node i to node j and 0 otherwise. For weighted networks, A_{ij} is generalized to represent the weight of the edge from i to j.

To investigate the structural regularities in a network, we suppose that the n nodes of the network fall into c groups whose memberships are unknown, i.e., we cannot observe or measure them directly. In this paper, we propose a statistical model to infer the group membership from the observed network structure.

The model we used is a kind of stochastic block model. A block model is a generative model and has a long tradition of study in the fields of social science and computer science. For a standard block model, a $c \times c$ matrix ω is generally adopted such that the matrix element ω_{rs} denotes the probability that a randomly selected edge connects group r to group s, i.e., the tail node of the edge is from group r and the head node is from s. The advantage of a block model is that the matrix ω explicitly characterizes various types of connecting patterns among groups.

In the standard block model, the nodes in the same group are identical, i.e., each node in a group has an equal probability to be the end node of an edge adjacent to the group. This constraint is relaxed in our model. Specifically, for an edge with its tail node from group r and its head node from group s, we use θ_{ri} to denote the probability that the tail node is i and ϕ_{sj} to denote the probability that the head node is j, respectively. In addition, we use $\overrightarrow{g}_{ij}$ and $\overleftarrow{g}_{ij}$ to denote, respectively, the group membership of the tail node and head node of the edge e_{ij}.

Until now, we have given all the quantities in our model. They can be classified into three classes: observed quantities $\{A_{ij}\}$, hidden quantities $\{\overrightarrow{g}_{ij}, \overleftarrow{g}_{ij}\}$, and model parameters $\{\omega_{rs},\theta_{ri},\phi_{sj}\}$. To simplify the notations, we henceforth denote A as the entire set $\{A_{ij}\}$, and, similarly, $\overrightarrow{g}$, $\overleftarrow{g}$, ω, θ, and ϕ for $\{\overrightarrow{g}_{ij}\}$, $\{\overleftarrow{g}_{ij}\}$, $\{\omega_{rs}\}$, $\{\theta_{ri}\}$, and $\{\phi_{sj}\}$, respectively.

With our model, an edge e_{ij} is generated in the following process:

(i) Select two groups $\overrightarrow{g_{ij}} = r$ and $\overleftarrow{g_{ij}} = s$, respectively, for the tail node and head node of the edge with probability ω_{rs}.

(ii) Draw the tail node i from the group r with probability θ_{ri}.

(iii) Draw the head node j from the group s with probability ϕ_{sj}.

Summing over the latent quantities r and s, the probability that we observe an edge e_{ij} can be written as

$$\mathrm{Prob}(e_{ij}|\omega,\theta,\phi) = \sum_{rs} \omega_{rs}\theta_{ri}\phi_{sj}. \tag{1}$$

Then, the likelihood of the observed network with respect to our model is

$$\mathrm{Prob}(A|\omega,\theta,\phi) = \prod_{ij} \left(\sum_{rs} \omega_{rs}\theta_{ri}\phi_{sj} \right)^{A_{ij}}. \tag{2}$$

Note that the self-loop edges are allowed and the weight A_{ij} is taken as the number of multi-edges connecting node i to node j, as done in many existing models including, for instance, the widely studied configuration model [28].

Intuitively, the parameter θ_{ri} characterizes the centrality of node i in the group r from the perspective of outgoing edges, while ϕ_{sj} describes the centrality of node j in the group s from the perspective of incoming edges. Different from traditional block models, by differentiating these two kinds of centrality, our model can provide more flexibility to explore broad types of intrinsic structural regularities in

networks. Note that the parameters ω_{rs}, θ_{ri}, and ϕ_{sj} satisfy the normalization conditions

$$\sum_{r=1}^{c}\sum_{s=1}^{c}\omega_{rs}=1,\quad \sum_{i=1}^{n}\theta_{ri}=1,\quad \sum_{j=1}^{n}\phi_{sj}=1. \tag{3}$$

Now our task is to estimate the model parameters and to infer the unobserved quantities by fitting the model to the observed network data. The standard framework for such a task is likelihood maximization. Generally, one works not with the likelihood [Eq. (2)] itself, but with its logarithm (log likelihood),

$$\mathcal{L}=\ln\mathrm{Prob}(A|\omega,\theta,\phi)=\sum_{ij}A_{ij}\ln\left(\sum_{rs}\omega_{r,s}\theta_{ri}\phi_{sj}\right). \tag{4}$$

The maximum of the likelihood and its logarithm are in the same place since the logarithm is a monotonically increasing function.

Directly maximizing the log likelihood is difficult because of the inner sum over the unobserved quantities, $\overrightarrow{g}_{ij}=r$ and $\overleftarrow{g}_{ij}=s$. Using Jensen's inequality, the maximization of the log likelihood can be transformed into the maximization of the expected log likelihood,

$$\begin{aligned}\overline{\mathcal{L}}&=\sum_{\overrightarrow{g},\overleftarrow{g}}\mathrm{Prob}(\overrightarrow{g},\overleftarrow{g}|A,\omega,\theta,\phi)\ln\mathrm{Prob}(A|\overrightarrow{g},\overleftarrow{g},\omega,\theta,\phi)\\&=\sum_{ijrs}\mathrm{Prob}(\overrightarrow{g}_{ij}=r,\overleftarrow{g}_{ij}=s|e_{ij},\omega,\theta,\phi)\\&\quad\times[A_{ij}(\ln\omega_{rs}+\ln\theta_{ri}+\ln\phi_{sj})]\\&=\sum_{ijrs}q_{ijrs}A_{ij}(\ln\omega_{rs}+\ln\theta_{ri}+\ln\phi_{sj}),\end{aligned} \tag{5}$$

where to simplify the notation we have defined $q_{ijrs}=\mathrm{Prob}(\overrightarrow{g}_{ij}=r,\overleftarrow{g}_{ij}=s|e_{ij},\omega,\theta,\phi)$, which denotes the probability that one observes an edge e_{ij} with its tail node i from group r and its head node j from group s, given the observed network and the model parameters.

With the expected log likelihood, we can give the best estimate of the value $\overline{\mathcal{L}}$, and the position of its maximum represents the best estimate of the most likely values of the model parameters. Specifically, if the value of q_{ijrs} is known, we can find the values of the model parameters ω, θ, and ϕ where $\overline{\mathcal{L}}$ reaches its maximum. However, the calculation of q_{ijrs} requires the values of these model parameters. To address such a problem, an expectation-maximization (EM) algorithm is adopted.

Under the framework of the EM algorithm, we first calculate the value of q_{ijrs} by

$$\begin{aligned}q_{ijrs}&=\frac{\mathrm{Prob}(\overrightarrow{g}_{ij}=r,\overleftarrow{g}_{ij}=s,e_{ij}|\omega,\theta,\phi)}{\mathrm{Prob}(e_{ij}|\omega,\theta,\phi)}\\&=\frac{\omega_{rs}\theta_{ri}\phi_{sj}}{\sum_{rs}\omega_{rs}\theta_{ri}\phi_{sj}}.\end{aligned} \tag{6}$$

Once we have the values of the q_{ijrs}, we can use them to evaluate the expected log likelihood and hence to find the values of ω, θ, and ϕ that maximize it.

Introducing the Lagrange multipliers ρ, γ_r, and λ_s to incorporate the normalization conditions in Eq. (3), the expected log-likelihood expression to be maximized becomes

$$\begin{aligned}\tilde{\mathcal{L}}=\overline{\mathcal{L}}&+\rho\left(1-\sum_{rs}\omega_{rs}\right)+\sum_{r}\gamma_r\left(1-\sum_{i}\theta_{ri}\right)\\&+\sum_{s}\lambda_s\left(1-\sum_{j}\phi_{sj}\right).\end{aligned} \tag{7}$$

By letting the derivative of $\tilde{\mathcal{L}}$ be 0, the maximum of the expected log likelihood occurs at the places where

$$\begin{aligned}\omega_{rs}&=\frac{\sum_{ij}A_{ij}q_{ijrs}}{\sum_{ijrs}A_{ij}q_{ijrs}},\\\theta_{ri}&=\frac{\sum_{js}A_{ij}q_{ijrs}}{\sum_{ijs}A_{ij}q_{ijrs}},\\\phi_{sj}&=\frac{\sum_{ir}A_{ij}q_{ijrs}}{\sum_{ijr}A_{ij}q_{ijrs}}.\end{aligned} \tag{8}$$

Equations (6) and (8) constitute our expectation-maximization algorithm. In the expectation step, the expected value of log likelihood is calculated through evaluating the values of q_{ijrs} with Eq. (6). In the maximization step, the expected value of log likelihood is maximized when the values of model parameters ω, θ, and ϕ are evaluated with Eq. (8). Implementation of the algorithm consists merely of iterating Eqs. (6) and (8) until convergence.

When the algorithm converges, we obtain a set of values for hidden quantity q_{ijrs} and model parameters ω, θ, and ϕ. This set of values is self-consistent with respect to Eqs. (6) and (8). However, it is not always the place where the log likelihood reaches its maximum. In other words, the expectation-maximization algorithm may converge to local maxima of the log likelihood. With different starting values, the algorithm will give rise to different solutions. To obtain a satisfactory solution, it is necessary to perform many runs with different starting values of model parameters and take the solution giving the highest log likelihood over all the runs performed.

By fitting the model to the observed network structure with the expectation-maximization algorithm, the estimated model parameters provide us with vital information for structural regularities of the network. Specifically, θ and ϕ describe the centrality of a node in groups containing it from the perspective of outgoing edges and incoming edges, respectively. The parameter ω characterizes the connecting patterns among different groups, i.e., the type of structural regularities.

More importantly, according to the model parameters, we can define two kinds of group memberships, α_{ir} and β_{js}, from the perspective of outgoing edges and incoming edges, respectively. Specifically, α_{ir} is the probability that node i is from group r when it acts as the tail node of the edges, while β_{js} is the probability that node j is from group s when it acts as the head node of the edges. For α_{ir}, it can be calculated by

$$\alpha_{ir}=\frac{\sum_{s}\omega_{rs}\theta_{ri}}{\sum_{rs}\omega_{rs}\theta_{ri}}. \tag{9}$$

Actually, α_{ir} provides a soft or fuzzy membership, i.e., node i can belong to more than one group simultaneously. When the identified structural regularity corresponds to community structure, we actually obtain the overlapping community structure, which has attracted much research attention ever since it was proposed. If one wants to get a hard partition, we can simply assign each node i to the group r satisfying $r = \arg\max_s\{\alpha_{is},\ s = 1,2,\ldots,c\}$. These statements for α_{ri} also apply to β_{ir} defined as

$$\beta_{js} = \frac{\sum_r \omega_{rs}\phi_{sj}}{\sum_{rs} \omega_{rs}\phi_{sj}}. \tag{10}$$

Finally, the model described above so far is based on directed networks. Actually, the model can be easily generalized to undirected networks by letting the parameter θ be identical to ϕ. The derivation follows the case of directed networks and the results are the same as Eqs. (6) and (8).

Now we discuss the computational cost of the expectation-maximization algorithm for the fitting of our model. For each iteration in this algorithm, the cost consists of two parts. The first part is from the calculation of q_{ijrs} using Eq. (6), whose time complexity is $O(m \times c^2)$. Here, m is the number of edges in the network and c is the number of groups. The second part is from the estimation of the model parameters using Eq. (8), whose time complexity is also $O(m \times c^2)$. We use T to denote the average number of iterations before the iteration process converges. Then, the total cost of the expectation-maximization algorithm for our model is $O(K \times T \times m \times c^2)$. Here, K is the number of times that the iteration process is restarted with different starting values to obtain a satisfactory solution. It is difficult to give a theoretical estimation for the number T of iterations. Generally speaking, T is determined by the network structure and the starting values of the model parameters. The number of runs is dependent on the scale of the network and its structural characteristics. For the networks tested in this paper, only less than 10 runs are needed to obtain a satisfactory result.

The computational cost limits our model to dealing with networks with tens of thousands of nodes. We look forward to seeing more efficient implementation for our model. Note that the method proposed in [29] provides a promising way to improve the computational efficiency and to decrease the memory space required. Finally, to make it convenient to evaluate the results in this paper and apply our model to more real world networks, we make the source computer code of our model available as Supplemental Material [30].

III. COMPARISON WITH OTHER MODELS

In this section, we illustrate the difference and connections between our model and several existing models. Figure 1 gives the schematic for our model and two existing generative models, namely, Newman's mixture model and Ren's probabilistic model.

For Newman's model, as shown in Fig. 1(a), each group r is characterized by the connecting preference θ_{rj} to node j, whether or not the node j is contained by the group r. The nodes belonging to the same group have a similar connecting preference. As a result, both assortative and disassortative

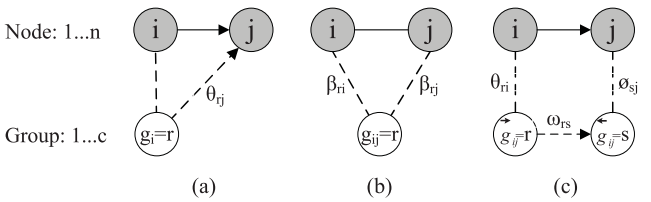


FIG. 1. Generative models for network data: (a) Newman's mixture model [24], (b) probabilistic model proposed in [22], and (c) our model. Filled circles represent observed quantities and unfilled ones correspond to hidden quantities. The solid line (with arrow) between node i and j indicates the existence of one (directed) edge connecting them. The dashed line connecting two circles indicates that the relation between the corresponding quantities is unobserved and must be learned from the observed network data. Arrows represent the directions of relation.

structural regularities can be detected by this model. However, this model has no parameter to explicitly characterize the type of the identified structure. More importantly, this model may produce a result that is a mixture of several types of structure and thus, in these cases, the identified structure may provide confusing information about the structural regularities. For example, for the network shown in Fig. 2, nodes 12, 15, 16, 19, 21, and 23 are identified by this model as overlapped nodes shared by the two groups, denoted by circles and squares, although these nodes only have connections to one of the two groups.

For Ren's model, as shown in Fig. 1(b), the two end nodes of each edge are assumed to be from the same group. As a result, only the assortative structure (community structure) can be detected using this model. Note that for this model, no edge is allowed to connect different groups. The relationship between communities is reflected by the overlapped nodes.

For our model, it essentially is a kind of stochastic block model, in which the relationships among different node groups are explicitly modeled by the block matrix w. In this way, our model possesses the advantages of both Newman's model and

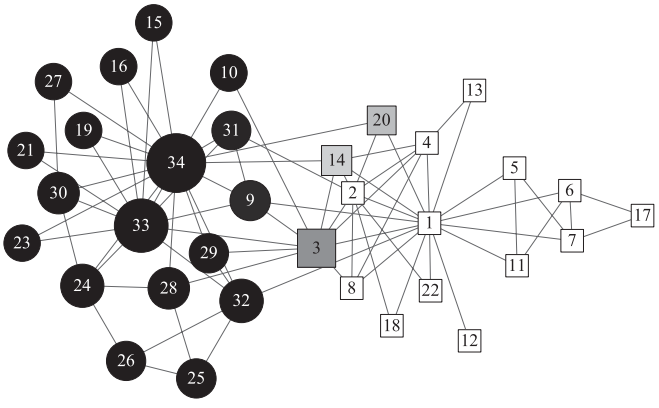


FIG. 2. The network of the karate club studied by Zachary [31]. The real social fission of this network is represented by two different shapes, i.e., circles and squares. The shades of the nodes indicate the mixed membership obtained by fitting our model to this network. The sizes of the nodes indicate the centrality degree (i.e., θ_{ri}) of nodes with respect to the left group. Here, θ_{ri} ranges from 0 for the smallest nodes to 0.22 for the largest nodes.

Ren's model, and overcomes the shortcoming of these two models.

On one hand, through learning the matrix w according to observed network data, various types of structural regularities can be explored by our model. The type of the identified structure is indicated by the matrix w. Specifically, when the matrix ω is an identity matrix, the identified structural regularity corresponds to an obvious community structure. Meanwhile, multipartite or anticommunity structure is revealed when the estimated model parameter ω is an antidiagonal matrix, with all of the antidiagonal elements being 1. For other types of structure, such as core-periphery structure and hierarchical structure, the form of ω is the same as the block matrix ω in traditional block models [26].

On the other hand, using the matrix w, our model discards the assumption of Ren's model that two end nodes of one edge are required to be from the same community. In this sense, Ren's model is a special case of our model. In addition, our model also provides several other flexibilities. By representing the centrality of nodes in a group from two different perspectives, i.e., according to the outgoing edges and incoming edges, our model can detect a broader range of structural regularities, which is beyond the capability of other models. This will be shown in the subsequent section. Moreover, our model can be further generalized by not requiring that the matrix w be a square matrix.

Finally, we compare our model to two recently proposed stochastic models for community detection [26,29]. First, both our model and Karrer's model [26] are stochastic block models, where a block matrix is adopted to characterize the connecting patterns among groups. The main difference between these two models lies in that Karrer's model is designed to detect disjoint structural regularities, while our model is for fuzzy structural regularities. This difference is reflected by the definition of the model parameters θ and ϕ in our model and the definition of the model parameter θ in Karrer's model. In addition, our model differentiates the outgoing edges from the incoming edges of the nodes, while Karrer's model does not. Second, similar to Ren's model, Ball's model [29] focuses on the community structure, while our model can uncover multiple types of structural regularities.

IV. EXPERIMENTAL RESULTS

In this section, we demonstrate the effectiveness of our model at exploring the structural regularities of networks by experiments on several real world or artificial networks with various types of intrinsic structural regularities. Then, we discuss the model selection issue, i.e., how to determine the optimal number of groups.

A. Detecting community structure

The test network is the famous karate club network constructed by Zachary. This network characterizes the acquaintance relationship between 34 members of a karate club at an American university. A dispute arose between the club's administrator and its principal karate teacher, and as a result the club eventually split into two smaller clubs, centered around the administrator and the teacher, respectively. The network and its fission are depicted in Fig. 2. The administrator and the teacher are represented by nodes 1 and 33, respectively.

TABLE I. Mixed membership of overlapped nodes.

Node	α_{i1}	q_{i1} [a]	$\frac{u_{1i}}{u_{1i}+u_{2i}}$ [b]
3	0.49	0.00	0.49
9	0.70	0.96	0.70
14	0.24	0.00	0.24
20	0.33	0.13	0.33
31	0.71	0.92	0.71
32	0.83	1.00	0.83

[a] q_{i1} is defined in [24] as the probability that node i belongs to group 1.
[b] $\frac{u_{1i}}{u_{1i}+u_{2i}}$ is defined in [22] as the probability that node i belongs to group 1.

By setting the group number $c = 2$, we fit our model to the karate club network data. The resulting matrix ω is a 2×2 identity matrix, indicating that the obtained structure is a community structure. Figure 2 shows the two groups found by our model with the expectation-maximization method. As shown in Fig. 2, the shades of the nodes in the figure represent the values of α_{i1},[1] where group 1 is the left group. As we can see, our model assigns most of the nodes strongly to one group or the other. Actually, all but six nodes are assigned 100% to one of the groups (black and white nodes in the figure). If we simply divide the nodes into two disjoint groups by assigning each node i to the group r according to the belong coefficients α_{ir}, the resulting groups perfectly correspond to the real split of the club.

In addition, Table I gives the belonging coefficient of the six overlapped nodes that are shared by the two groups. These overlapped nodes are nodes 3, 9, 14, 20, 31, and 32. Note that these overlapped nodes are often misclassified by traditional partition-based community detection methods. For comparison, we also give the mixed membership of these six nodes according to Newman's mixture model and Ren's model. As we can see, our model and Ren's model produce the same results, which is attributed to the fact that Ren's model is a special case of our model. However, Newman's model behaves very differently from the other two models. Actually, for Newman's model, another 10 nodes are also assigned to both of the two groups, e.g., nodes 12 and 15. Such a result is counterintuitive to the real structure of this network. In conclusion, our model performs better than Newman's model at detecting the overlaps between groups. Ren's model can only detect community structure, while our model can detect other types of structural regularities, as illustrated in the following test.

B. Detecting multipartite structure

Now we illustrate the detection of multipartite or anticommunity structure according to our model. The test network is the adjacency network of English words taken from Ref. [9]. In this network, the nodes represent 112 commonly occurring

[1] Since this network is an undirected network, the two kinds of belonging coefficient are identical, i.e., $\alpha_{ir} = \beta_{ir}$.

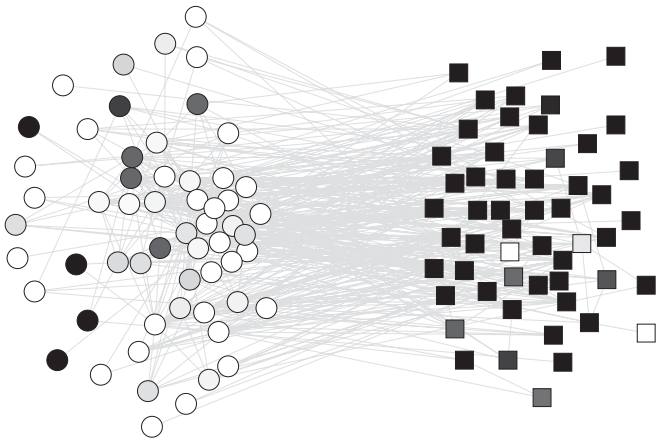


FIG. 3. The adjacency network of English words. Node groups corresponding to adjectives and nouns are denoted by circles and squares, respectively. The shades of nodes indicate their belonging coefficient obtained by fitting our model to this network.

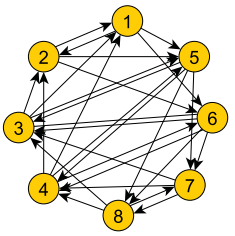
adjectives and nouns in the novel *David Copperfield* by Charles Dickens, with edges connecting any pairs of words that appear adjacent to each other at any place in the text. Generally, adjectives occur next to nouns in English. Thus, most edges in the network connect an adjective to a noun and the network is approximately bipartite, i.e., this network possesses anticommunity structure. This can be seen clearly in Fig. 3, where the adjectives and nouns are represented by circles and squares, respectively.

Fitting our model to this network with $c = 2$, the resulting ω is a 2×2 antidiagonal matrix, indicating that the identified structure is a bipartite structure. The obtained two groups and node memberships are shown by the shades of nodes in Fig. 3. We can see that most nodes are assigned to only one group, although there are several ambiguous cases corresponding to the nodes with intermediate shades. If we assign each node to its most preferred group, the resulting two disjoint groups well separate the adjectives from the nouns. In fact, 100 of the 112 total nodes are correctly classified. This accuracy is the same as the result given by Newman's mixture model.

As a comparison, we also apply Ren's model to this network by setting the group number to 2. Only 60 nodes of the 112 total nodes are correctly classified, similar to the accuracy of random assignment. The ineffectiveness of Ren's model in this network is attributed to the fact that Ren's model presumes the existence of community structure in the network, while the intrinsic structural regularity is a bipartite structure.

C. Exploring other types of structural regularity

In the previous tests, we have demonstrated that our model can be used to detect both the assortative structure (i.e., community structure) and the disassortative structure (i.e., multipartite structure) without knowledge of which type of structural regularities exists in the target networks. Now we will further show that our model can also detect other types of structure, which cannot be revealed by competing models.



(a)

Out \ In	1-2 5-6	3-4 7-8
1-4	Yes	No
5-8	No	Yes

(b)

FIG. 4. (Color online) (a) A schematic network. The directed edges are placed according to the rules described in (b).

We consider the schematic network depicted in Fig. 4(a). This network is constructed according to the rules in Fig. 4(b). Intuitively, according to the outgoing edges in this network, the nodes can be divided into two groups: {1,2,3,4} and {5,6,7,8}. Meanwhile, according to the incoming edges, the nodes of this network belong to another two groups: {1,2,5,6} and {3,4,7,8}.

We apply Newman's model, Ren's model, and our model to this schematic network. Limited by the assumptions of models, both Newman's model and Ren's model fail to uncover the intrinsic structural regularity indicated by the construction rules. For our model, the flexibility of model assumption enables it to accurately detect this type of structure. Specifically, by fitting our model to this network, the obtained θ or α reveals the two groups indicated by the outgoing edges, while the ϕ or β reflects the two groups indicated by the incoming edges.

D. Model selection issue

In the previous tests, we needed to specify the group number before fitting our model to a network. However, the group number is unknown *a priori* for many cases. Thus, it is helpful to give a criterion to determine the appropriate group number for a given network. This task is known as the model selection issue in statistics. We deal with this problem by using the minimum description length principle, which is also used to handle the model selection issue in Ren's model.

According to the minimum description length principle, the required length to describe the network data is composed of two parts. The first part describes the coding length of the network using our model. This coding length is $-L$ for a directed network and $-L/2$ for an undirected network. The second part gives the length for coding model parameters. This part is $-\sum_{rs} \ln\omega_{rs} - \sum_{ri}(\ln\theta_{ri} + \ln\phi_{ri})$ for a directed network and $-\sum_{rs} \ln\omega_{rs} - \sum_{ri} \ln\theta_{ri}$ for an undirected network. In this way, the optimal c is the one that minimizes the total description length.

As tests, we consider two real world networks with prior knowledge of the intrinsic group numbers. These two networks are, respectively, the journal citation network constructed in Ref. [32] and the American football team network described in Ref. [1]. In the journal citation network, each node corresponds to a journal, in which all 40 journals are from four different fields: multidisciplinary physics, chemistry, biology, and ecology. Journals from the same field are more likely connected by citation relation. For the football network,

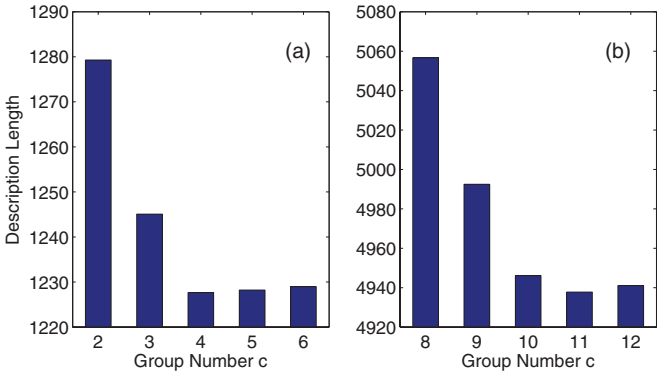


FIG. 5. (Color online) Model selection results for the (a) journal citation network and (b) American football team network.

nodes represent the 115 teams, respectively, belonging to 12 conferences, and generally games are more frequent between teams who belong to the same conference than between teams of different conferences.

As shown in Fig. 5, the number of intrinsic groups is correctly identified for the journal citation network. However, for the football network, 11 is the optimal number of groups while the intrinsic number is 12. By checking the found node groups, we find that only 11 node groups have their identities, i.e., each group contains at least one node after assigning nodes to their most preferred groups according to the obtained belonging coefficient α or β. This indicates that the appropriate group number is 11 for the football network. In fact, many well-known community detection methods also identify 11 communities.

V. CONCLUSIONS

In this paper, we have studied the exploration of intrinsic structural regularities in networks using a general stochastic block model. Without prior knowledge, our model not only can detect broad types of intrinsic structural regularities, but also can learn the type of identified structure directly from the network data. Tests on a number of artificial and real world networks demonstrate that our model outperforms the state-of-the-art models at shedding light on the structural features of networks. This flexibility enables our model to be an effective way to reveal the structural regularities of networks and further to help us understand the relationship between the structure and function of networks. For potential applications, our model can be used to predict the emergence or vanishing of edges in networks. In a future work, we will generalize our model by eliminating the requirement that the block matrix be a square matrix, and we will investigate the possible applications of the more flexible model.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China under Grants No. 60873245, No. 60933005, No. 60873217, and No. 60803123. This work was also partly funded by the National High-Tech R&D Program of China (863 Program) with Grant No. 2010AA012502. The authors thank Alex Arenas, Mark Newman, and Santo Fortunato for providing network and other data used in this paper. The authors also thank Fu-Xin Ren, Jun-Ming Huang, and Lu Bai for helpful discussions and useful suggestions. Finally, the authors acknowledge the anonymous reviewers for valuable comments on this paper.

[1] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, IEEE Comput. **35**, 66 (2002).
[4] X. Q. Cheng, F. X. Ren, S. Zhou, and M. B. Hu, New J. Phys. **11**, 033019 (2009).
[5] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[6] X. Q. Cheng, F. X. Ren, H. W. Shen, Z. K. Zhang, and T. Zhou, J. Stat. Mech.: Theory Exp. (2010) P10011.
[7] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[8] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).
[9] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[10] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[11] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. USA **101**, 2658 (2004).
[12] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[13] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[14] H. W. Shen, X. Q. Cheng, K. Cai, and M. B. Hu, Physica A **388**, 1706 (2009).
[15] H. W. Shen, X. Q. Cheng, and J. F. Guo, J. Stat. Mech.: Theory Exp. (2009) P07042.
[16] X. Q. Cheng and H. W. Shen, J. Stat. Mech.: Theory Exp. (2010) P04024.
[17] H. W. Shen, X. Q. Cheng, and B. X. Fang, Phys. Rev. E **82**, 016114 (2010).
[18] H. W. Shen and X. Q. Cheng, J. Stat. Mech.: Theory Exp. (2010) P10020.
[19] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[20] J. J. Ramasco and M. Mungan, Phys. Rev. E **77**, 036122 (2008).
[21] A. Vazquez, Phys. Rev. E **77**, 066106 (2008).
[22] W. Ren, G. Y. Yan, X. P. Liao, and L. Xiao, Phys. Rev. E **79**, 036111 (2009).
[23] H. Z. Zhang, B. J. Qiu, C. L. Giles, H. C. Foley, and J. Yen, IEEE Intell. Secur. Inform. 200 (2007).
[24] M. E. J. Newman and E. A. Leicht, Proc. Natl. Acad. Sci. USA **104**, 9564 (2007).
[25] A. Clauset, C. Moore, and M. E. J. Newman, Nature (London) **453**, 98 (2008).

[26] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).
[27] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, J. Mach. Learn. Res. **9**, 1981 (2008).
[28] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[29] B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **84**, 036103 (2011).
[30] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.84.056111 for the C++ implementation using the expectation-maximization algorithm for our GSB model. See also http://searchforum.org.cn/research/shw/homepage.html.
[31] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[32] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. USA **104**, 7327 (2007).