Electronic Journal of Statistics
Vol. 18 (2024) 4724–4766
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2309

Subsampling-based modified Bayesian information criterion for large-scale stochastic block models

Jiayi Deng[1], **Danyang Huang**[*2], **Xiangyu Chang**[3] **and Bo Zhang**[*2]

[1]*Department of Statistics and Epidemiology, Graduate School of the PLA General Hospital, Beijing, China*

[2]*Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China,*
e-mail: dyhuang@ruc.edu.cn; mabzhang@ruc.edu.cn

[3]*School of Management, Xi'an Jiaotong University, Xi'an, China*

Abstract: Identifying the number of communities is a fundamental problem in community detection, which has received increasing attention recently. However, rapid advances in technology have led to the emergence of large-scale networks in various disciplines, thereby making existing methods computationally infeasible. To address this challenge, we propose a novel subsampling-based modified Bayesian information criterion (SM-BIC) for identifying the number of communities in a network generated via the stochastic block model and degree-corrected stochastic block model. We first propose a node-pair subsampling method to extract an informative subnetwork from the entire network, and then we derive a purely data-driven criterion to identify the number of communities for the subnetwork. In this way, the SM-BIC can identify the number of communities based on the subsampled network instead of the entire dataset. This leads to important computational advantages over existing methods. We theoretically investigate the computational complexity and identification consistency of the SM-BIC. Furthermore, the advantages of the SM-BIC are demonstrated by extensive numerical studies.

Keywords and phrases: Network community detection, large-scale networks, network subsampling, model selection.

Received December 2022.

Contents

*Corresponding author

1. Introduction

Network community detection is one of the most widely-studied topics in network analysis [31, 58, 29]. Intuitively, for networks with assortative communities, community detection aims to distribute the network nodes to several clusters, so that nodes in the same cluster have denser connectivity [4]. Network community structure is beneficial for understanding the characteristics of each cluster [31, 9]. Specifically, in social network platforms (e.g., *Facebook*, *Twitter*, and *Sina Weibo*), communities can be formed by users with similar interests or preferences, which enables online platforms to recommend suitable products and services to targeted groups [11, 6, 67].

In the past few decades, numerous assortative community detection methods have been proposed, including but not limited to modularity maximization [57, 34], spectral clustering [59, 73, 65], belief propagation [37, 81], and pseudo-likelihood methods [3, 77]. Theoretically, the stochastic block model (SBM), has been widely assumed to analyze the consistency properties of network community methods [39, 69, 60]. It should be noted that most community detection methods require the number of communities K_0 to be known in advance. Then, the theoretical properties can be carefully established. However, K_0 is typically unknown in real-world networks. Therefore, how to choose K_0 is important.

A variety of methods have been proposed to determine the number of communities K_0, such as the eigenvalue-based methods [47, 12, 10], semi-definite programming-based methods [52, 79], network cross-validation methods [18, 51], and likelihood-based methods [22, 78, 41, 54]. Specifically, the eigenvalue-based methods estimate the number of communities based on the eigenvalue properties of non-backtracking, Bethe Hessian, or normalized Laplacian matrices [47, 12, 10, 21, 42]. Additionally, the semi-definite programming approach identifies K_0 by solving a semi-definite optimization problem [52, 79]. Moreover, the network cross-validation method extends the cross-validation method to network data via a network sampling strategy [18, 51]. The likelihood-based approaches aim to make full use of observed samples, which have been widely studied, including Bayesian information criterion and likelihood ratio methods. Specifically, the Bayesian information criterion consists of a conditional log-likelihood of entire observations and a penalty term that depends on the prior distribution of the latent variable [22, 66, 41, 14]. The likelihood ratio approaches are based on a stepwise goodness-of-fit estimator to determine the number of communities in networks [78, 54, 44]. It is remarkable that to evaluate each candidate K via the aforementioned criteria, such as the network cross-validation methods and the likelihood-based approaches, we need to first estimate the parameters for the SBM using the entire observed network. In this case, spectral clustering is considered a simple and easy-to-implement approach with well-founded theoretical guarantees [65, 16, 86, 49].

However, recent advances in science and technology have brought about large-scale network data, leading to unprecedented computational challenges for community detection. For example, as reported by *Statista* (*www.statista.com*), in January 2022, the online social networks *Facebook*, *Twitter*, and *Sina Weibo* had approximately 2,910 million, 436 million, and 573 million active users, respectively. Researchers could also access the relationships of millions of network nodes using open-source datasets, such as the *Stanford Large Network Dataset*[1], which has collected different networks with more than 10 million nodes each. Consequently, directly applying traditional methods to estimate K_0 for these large-scale network data is impractical. For example, for a network with N nodes, the time complexity of spectral clustering-based methods is no lower than $O(N^3)$ for estimating K_0 [80, 50, 20]. Even if the algorithm could be accelerated, the computational complexity is still in the order of $O(N^2)$ [36, 28, 55]. To deal with the computational challenge brought by large-scale networks, subsampling is a valuable tool [62]. Its main advantage is that we can obtain a computationally efficient and consistent estimator based on a small subsample [76, 75, 74, 82]. Although subsampling pays the price of statistical convergence, it makes the traditional methods feasible in large-scale data analysis.

In the literature, various sampling designs have been proposed to derive representative samples of a given network, which include node sampling methods [68, 8, 56, 40] and edge sampling methods [32, 33, 51]. The node sampling methods select landmark nodes from the entire network, and the subnetwork

[1] http://snap.stanford.edu

is induced by these selected nodes. Uniform node sampling is considered to be the simplest method and has been widely used [68, 8, 53, 56]. A few studies investigate the statistical inference about key network characteristics through node sampling [35, 84, 7]. Additionally, the node sampling method has been investigated in network community detection [24, 15]. Another widely studied node sampling method is snowball sampling [43, 61, 19]. Based on the snowball sampling approach, [70] and [2] recently developed bootstrap methods to reduce estimation bias for large networks.

The edge sampling methods randomly collect edge samples from the entire network, which have also received considerable attention [27, 26, 51]. For example, [27] and [32] employed edge sampling procedures to estimate the average degree of a network, while [25] studied the network degree distribution using the network bootstrap method. Recently, edge sampling approaches have been investigated to approximate counting the number of subgraphs [33, 26, 5]. Moreover, [51] applied uniform edge sampling in random graph model selection. Note that existing studies focus on subsampling many times to provide stable statistical inference for network models. However, we aim to conduct subsampling only once to allow model selection for large-scale networks with limited computational resources.

This work proposes a novel subsampling-based modified Bayesian information criterion (SM-BIC) for identifying the number of communities for large-scale SBMs. Specifically, in the context of large-scale networks, we first develop a *node-pair subsampling* method to extract a subnetwork from the entire network. The node-pair subsampling method combines the idea of uniform node sampling and edge sampling. More precisely, we first uniformly and randomly select a subset of nodes from the entire network and then collect all edges related to these nodes to construct a subnetwork. In this way, this subnetwork fully retains the connection information between the selected nodes and the entire network. Note that the node-pair subsampling method only requires subsampling once due to computational efficiency. Then, based on the selected subnetwork, we derive a purely data-driven criterion without tuning any parameters. Since the criterion is based only on subsampled data, it makes the subsequent parameter estimation applicable even for large-scale networks with affordable computational resources. In particular, we use spectral clustering for the subsampled subnetwork to obtain the community assignments. In this way, the computational complexity of the SM-BIC can be as low as $O(Nn)$, where n is the subsample size satisfying $n << N$. Furthermore, we extend the SM-BIC to the degree-corrected stochastic block model (DCSBM) [45]. We theoretically investigate the computational advantage of the SM-BIC. Most importantly, for both the SBM and DCSBM, we establish the consistency of the SM-BIC by studying the penalized log-likelihood function under misspecification cases (e.g., under-fitting and over-fitting).

1.1. Contributions of the paper

The advantages of the proposed SM-BIC method are listed as follows. First, compared with the eigenvalue-based methods [47, 12, 10, 21, 42], the SM-BIC

fully exploits the connectivity information in the selected subnetwork, while the eigenvalue-based methods use the eigenvalue information of network matrices. Second, compared with the method based on semi-definite programming [52, 79], the proposed SM-BIC method applies the spectral clustering algorithm to identify community labels for network nodes, which is more computationally efficient. Third, compared with the network cross-validation methods [18, 51], the SM-BIC only requires subsampling once, while the network cross-validation method uses a network resampling technique, which requires tuning the number of folds.

Additionally, compared with the aforementioned BIC-based approaches [22, 66, 41, 14] and likelihood ratio methods [78, 54, 44], the SM-BIC can identify K_0 using only a small subnetwork; further, it is a completely data-driven method without any predefined tuning parameters. Consequently, the SM-BIC could be feasibly applied to identify the number of communities for large-scale networks with affordable computational resources. Specifically, its computational complexity could be as low as $O\{N(\log N)^2\}$, as demonstrated in Propositions 1 and 2.

The remainder of this paper is organized as follows. In Section 2, we introduce the subsampling-based modified Bayesian information criterion. In Section 3, we discuss the theoretical properties of the SM-BIC and establish the consistency of the estimator of the number of communities. In Section 4, we demonstrate the effectiveness of our method through extensive numerical studies. Further discussions are provided in Section 5. Proofs are presented in the Appendices and the supplementary materials.

2. Subsampling-based modified Bayesian information criterion for stochastic block model

In this section, we first introduce the stochastic block model and challenges of existing model selection methods. Then, we develop the SM-BIC for large-scale SBMs and extend the criterion to DCSBMs. Lastly, we discuss the parameter estimation procedure for this method.

2.1. Preliminaries

Consider a large-scale undirected graph generated from an SBM with N nodes and K_0 communities. The observed random graph is often represented by a symmetric adjacency matrix $A \in \mathbb{R}^{N\times N}$ with zero diagonal entries. Specifically, for any node pair (i, j), if there is a connection, then $A_{ij} = 1$; otherwise, $A_{ij} = 0$. For each node i, denote its community label as $g^*_{N,i} \in [K_0] = \{1, \ldots, K_0\}$. Let $N_{k,g^*_N} = \sum_i \mathbb{I}(g^*_{N,i} = k)$ denote the size of the k-th cluster. Given a label vector $g^*_N = (g^*_{N,1}, \ldots, g^*_{N,N})^\top \in [K_0]^N$, the edge variables A_{ij}s for $i < j$ are independent Bernoulli random variables with $\mathbb{E}(A_{ij}) = B^*_{g^*_{N,i}g^*_{N,j}}$, where $B^* = (B^*_{kl}) \in (0, 1)^{K_0\times K_0}$ is a symmetric matrix describing connectivity probability within and between communities. Namely, each element $B^*_{kl} \in (0, 1)$ represents

the connectivity probability between k and l communities ($1 \leq k, l \leq K_0$). In this way, the connectivity probability between any node pair (i, j) depends only on their community labels. For simplicity, let $\mathrm{SBM}_{K_0}(g_N^*, B^*)$ represent a stochastic block model with K_0 blocks parameterized by g_N^* and B^*.

Throughout this paper, we let g_N^* and B^* denote the true parameters of the observed adjacency matrix A. Furthermore, K_0 is considered to be a fixed constant. Under any candidate K, denote $g_N \in [K]^N$ as the community assignment of the K-block model, and the corresponding connectivity matrix is represented by a symmetric matrix $B \in \mathbb{B}_K = (0,1)^{K \times K}$. Additionally, when we refer to model selection, we mean the selection of K_0 for $\mathrm{SBM}_{K_0}(g_N^*, B^*)$.

For the likelihood-based methods, to determine the number of communities, it is necessary to estimate the community assignment g_N for each candidate K. For super-large N, even if accelerated algorithms are adopted, the computational cost is still high. For example, the randomized spectral clustering algorithm [83] has computational cost in the order of $O(N^2)$. This motivates us to develop a network subsampling-based model selection criterion that reduces the cost by investigating small subsamples.

2.2. Subsampling-based modified Bayesian information criterion

In the context of large-scale networks, we first introduce the network subsampling method. Note that, unlike independent data, network data are correlated with each other by connections. To characterize the community membership of network nodes, we use a node-pair subsampling method to collect a subnetwork from the entire network. Specifically, we first uniformly sample n nodes from $[N]$ without replacement; that is, the probability of each node being selected is equal to n/N, where the subsample size $n << N$. We further denote the set of selected nodes as $\mathcal{S} = \{s_j \in [N] : \text{node } s_j \text{ is selected}\}$. Then, we sample all node pairs related to these selected nodes. That is, if node i is selected and there is a connection between i and s_j, then node pair (i, s_j) is also collected. The subsampling method is illustrated in Figure 1. We refer to this method as *node-pair subsampling*. For convenience, let j ($j \in [n]$) denote the index of the selected node s_j in the node set $\mathcal{S}$. Define a $N \times n$ matrix $A^{\mathcal{S}}$ to represent these selected connections, where the entries are $A^{\mathcal{S}}_{ij} = A_{is_j}$, for $i \in [N], s_j \in \mathcal{S}$. Then, we focus on the observation $A^{\mathcal{S}}$ rather than the entire network connections, to identify the number of communities.

For model selection, we introduce the proposed modified Bayesian information criterion based on $A^{\mathcal{S}}$. The criterion is derived from the maximization of the log-posterior likelihood function of g_N. We first provide the prior distribution of g_N under SBM_K. Based on the selected sample $A^{\mathcal{S}}$, we demonstrate that the community partition of the entire network is determined by the community assignment of the selected nodes. Specifically, consider the community assignment of the selected nodes to be g_n ($g_n \in [K]^n$), and $g_{n,j}$ is the community label of the selected node s_j. Then, for any unselected node $i \notin \mathcal{S}$, we have different ways to obtain its community label based on the label of the se-

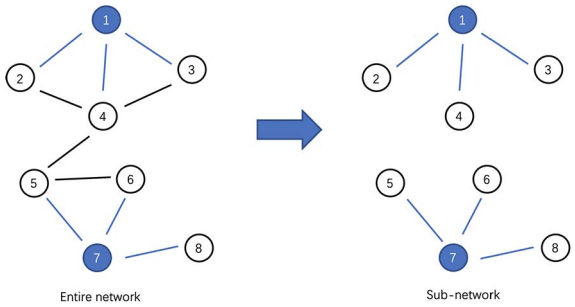

FIG 1. *An example of node-pair subsampling. The left panel shows the entire network, where the colored nodes are considered to be selected by simple random sampling without replacement, whereas their corresponding connections (shown in dark blue) are extracted from the entire network. The right panel represents the subnetwork obtained by the node-pair subsampling method.*

lected nodes. For example, we could cluster this node to the community with the most connections to it. Namely, the community label of the unselected node i is given by $\widehat{g}_{N,i} = \max_k \sum_{1\le j\le n} A^{\mathcal{S}}_{ij}\mathbb{I}(g_{n,j} = k)$, where $\mathbb{I}(\cdot)$ is an indicator function. We could alternatively obtain the label assignment for unselected nodes by spectral clustering, which is illustrated in detail in the next subsection. In this way, based on $A^{\mathcal{S}}$, the set of all possible community assignments for entire network nodes is provided as $\mathcal{C}(A^{\mathcal{S}}, K) = \bigcup_{g_n\in[K]^n} \{g_N \in [K]^N : \forall\, i \notin \mathcal{S}, g_{N,i} = \max_k \sum_{1\le j\le n} A^{\mathcal{S}}_{ij}\mathbb{I}(g_{n,j} = k), \forall\, s_j \in \mathcal{S}, g_{N,s_j} = g_{n,j}\}$. Therefore, the number of possible community assignments is $|\mathcal{C}(A^{\mathcal{S}}, K)| = K^n$. Similar to [17], we assign the prior probability to g_N as

$$\phi(g_N) = K^{-n}, \text{ for } g_N \in \mathcal{C}(A^{\mathcal{S}}, K). \tag{2.1}$$

Next, we analyze the posterior probability of g_N.

We start with studying the probability of $A^{\mathcal{S}}$ under SBM_K. We denote the set of node pairs corresponding to the independent edge variables in $A^{\mathcal{S}}$ by $E = E_{\text{in}} \cup E_{\text{out}}$. Where $E_{\text{in}} = \{(i, s_j) : i, s_j \in \mathcal{S}, s_j > i\}$ and $E_{\text{out}} = \{(i, s_j) : i \in [N] - \mathcal{S}, s_j \in \mathcal{S}\}$ represent the set of node pairs within selected nodes and that between selected and unselected nodes, respectively. Moreover, since $|E_{\text{in}}| = n(n-1)/2$ and $|E_{\text{out}}| = (N-n)n$, we have $|E| = Nn - n(n+1)/2$. Let $o_{kl,g_N} = \sum_{(i,s_j)\in E} A^{\mathcal{S}}_{ij}\mathbb{I}(g_{N,i} = k, g_{N,s_j} = l)$ and $n_{kl,g_N} = \sum_{(i,s_j)\in E}\mathbb{I}(g_{N,i} = k, g_{N,s_j} = l)$ denote the number of observed connections and the number of maximum possible connections between (k, l) clusters, respectively. Additionally, define a vector $\theta \in \Theta_K = (0, 1)^{K(K+1)/2}$ to represent the upper triangle elements of B. Then, given (g_N, θ), the log-likelihood function of $A^{\mathcal{S}}$ is

$$\log f(A^{\mathcal{S}}|g_N, \theta) = \sum_{1\le k\le l\le K} \{o_{kl,g_N} \log\theta_{kl} + (n_{kl,g_N} - o_{kl,g_N}) \log(1 - \theta_{kl})\}.$$

Accordingly, the likelihood function of g_N is given by, $f(A^{\mathcal{S}}|g_N) = \int f(A^{\mathcal{S}}|g_N, \theta)\, p(\theta)\mathrm{d}\theta$, where $p(\theta)$ is the prior distribution of θ.

Then, we give an approximation of the log-likelihood function $\log f(A^{\mathcal{S}}|g_N)$ in the following lemma.

Lemma 1 (Log-likelihood function approximation)**.** *Suppose the adjacency matrix A generated from SBM_K and the subset of nodes $\mathcal{S}$ collected by simple random sampling n nodes from the entire network without replacement. Then, based on subsampled adjacency matrix $A^{\mathcal{S}}$, the log-likelihood function $\log f(A^{\mathcal{S}}|g_N)$ can be approximated by,*

$$\log f(A^{\mathcal{S}}|g_N) = \sup_{\theta\in\Theta_K} \log f(A^{\mathcal{S}}|g_N,\theta) - \frac{K(K+1)}{4}\log M + O(1), \tag{2.2}$$

where M denotes the number of independent edge variables in $A^{\mathcal{S}}$, i.e., $M = |E| = Nn - n(n+1)/2$, and the term $O(1)$ is relevant to the asymptotic regime related to N.

The proof of Lemma 1 can be found in Appendix B.1. As a result, under SBM_K, according to (2.1) and (2.2), the log-posterior probability of g_N is

$$\log f(g_N|A^{\mathcal{S}}) = \log\{f(A^{\mathcal{S}}|g_N)\phi(g_N)\} + c, \tag{2.3}$$

where $c = -\int \log\{f(A^{\mathcal{S}}|g_N)\phi(g_N)\}\mathrm{d}g_N$ is a constant.

We now establish the SM-BIC. According to Bayesian inference, the community assignment that maximizes the posterior probability is estimated, that is $\widehat{g}_N = \max_{g_N\in\mathcal{C}(A^{\mathcal{S}},K)} \log f(g_N|A^{\mathcal{S}})$. To this end, based on (2.2) and (2.3), the SM-BIC is proposed as follows:

$$\ell(K) = \max_{g_N\in\mathcal{C}(A^{\mathcal{S}},K)} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N,B) - \left\{ n\log K + \frac{K(K+1)}{4}\log M \right\}. \tag{2.4}$$

The form of the criterion (2.4) seems to be similar to the corrected BIC criterion proposed by [41]. However, there are two key differences from the corrected BIC, which are also the key contributions of our criterion. First, the SM-BIC is a purely data-driven method without any predefined tuning parameters, whereas the corrected BIC requires choosing one parameter to control the model selection results. This is because we assume a simple uniform prior for SBM_K and the latent label vector g_N; this prior setting follows the work of [17]. Second, based on (2.4), we estimate the community assignment from $A^{\mathcal{S}}$, which has a lower dimension than A for $n << N$. Hence, criterion (2.4) could save computational costs. We demonstrate the important computational advantages of the SM-BIC in Subsection 2.4.

2.3. Extension to degree-corrected stochastic block model

The DCSBM [45] is generalized from the SBM, which introduces node-specific parameters to allow for degree heterogeneity within communities. Specifically, given parameters g_N, B, the probability of an edge between (i,j) is represented

by $P(A_{ij}=1)=\psi_i B_{g_{N,i}g_{N,j}}\psi_j$, where the parameter ψ_i characterizes the individual activeness of node i. In this way, a DCSBM is parameterized by a triplet (g_N, B, ψ) where $\psi=(\psi_1,\ldots,\psi_N)^\top$. For consistency, we assume that the underlying model is $\mathrm{DCSBM}_{K_0}(g_N^*, B^*, \psi^*)$. For identifiability of this model, the constraint $\sum_i \psi_i^* \mathbb{I}(g_{N,i}^*=k)=N_{k,g_N^*}$ is imposed on each community $1\le k\le K_0$. Then, we extend the SM-BIC to the DCSBM.

We start with the log-likelihood function of the subsampled adjacency matrix $A^{\mathcal{S}}$. Similar to [45] and [86], we replace Bernoulli likelihood with Poisson likelihood and assume $A_{ij}\sim \mathrm{Poisson}(\psi_i B_{g_{N,i}g_{N,j}}\psi_j)$ to simplify the derivation. Furthermore, let $n_{kl,g_N}(\psi)=\sum_{(i,s_j)\in E}\psi_i\psi_{s_j}\mathbb{I}(g_{N,i}=k, g_{N,s_j}=l)$. In this way, under $\mathrm{DCSBM}_K(g_N, B, \psi)$, the log-likelihood function of the subsampled adjacency matrix $A^{\mathcal{S}}$ is given by

$$\begin{aligned}&\log f(A^{\mathcal{S}}|g_N, B, \psi)\\ &\quad=\sum_{(i,s_j)\in E} A_{ij}^{\mathcal{S}}\log(\psi_i\psi_{s_j})+\sum_{1\le k\le l\le K}\{o_{kl,g_N}\log B_{kl}-n_{kl,g_N}(\psi)B_{kl}\}.\end{aligned}$$

Then, we consider ψ in two cases. First, if ψ is known, according to (2.4), the SM-BIC of the DCSBM is proposed as follows:

$$\ell(K)=\max_{g_N\in\mathcal{C}(A^{\mathcal{S}},K)}\sup_{B\in\mathbb{B}_K}\log f(A^{\mathcal{S}}|g_N,B,\psi)-\left\{n\log K+\frac{K(K+1)}{4}\log M\right\}. \tag{2.5}$$

Second, if ψ is unknown, we take a plug-in estimator $\widehat{\psi}$ into the (2.5) criterion to replace ψ. In this case, an estimation of ψ is provided in the following subsection.

2.4. Parameter estimation based on subsampled adjacency matrix

Here, we first introduce how to apply the SM-BIC to determine the number of communities for large-scale SBMs. Specifically, based on *node-pair* subsampling, we evaluate each candidate K through the following three steps: label assignment, parameter estimation, and SM-BIC calculation. Thereafter, we further present the estimation method of the degree heterogeneity cases.

Label assignment We first perform the label assignment step on the $N\times n$ subsampled adjacency matrix. For a candidate K and subsampled adjacency matrix $A^{\mathcal{S}}$, the extended spectral clustering algorithm can be accomplished as follows.

(1) Perform SVD on $A^{\mathcal{S}}$, and extract the largest K left eigenvectors, denoted as $V_1,\ldots,V_K$, and define a $N\times K$ matrix $V=(V_1,\ldots,V_K)$ to represent the embedding matrix.
(2) Apply K-means clustering to the rows of V to estimate node assignments and denote the clustering results by $\widehat{g}_N$.

Parameter estimation Based on the estimated label vector $\widehat{g}_N$, we construct the plug-in estimator for the connectivity matrix B. Specifically, for all $1 \le k \le l \le K$, the estimated (k,l)-th element of $\widehat{B}$ is

$$\widehat{B}_{kl} = \frac{o_{kl,\widehat{g}_N}}{n_{kl,\widehat{g}_N}} = \frac{\sum_{(i,s_j)\in E} A^{\mathcal{S}}_{ij}\mathbb{I}(\widehat{g}_{N,i}=k,\widehat{g}_{N,s_j}=l)}{\sum_{(i,s_j)\in E}\mathbb{I}(\widehat{g}_{N,i}=k,\widehat{g}_{N,s_j}=l)}, \tag{2.6}$$

and taking $\widehat{B}_{lk} = \widehat{B}_{kl}$, we obtain the estimated connectivity matrix $\widehat{B}$.

SM-BIC calculation Based on $(\widehat{g}_N, \widehat{B})$, we evaluate the estimated SBM_K $(\widehat{g}_N, \widehat{B})$ by

$$\widehat{\ell}(K) = \log f(A^{\mathcal{S}}|\widehat{g}_N, \widehat{B}) - \left\{ n\log K + \frac{K(K+1)}{4}\log M \right\}. \tag{2.7}$$

Therefore, we choose K which maximizes the SM-BIC (2.7) as the number of communities.

Algorithm 1 Model Selection Algorithm for SBM.

Input: adjacency matrix $A^{\mathcal{S}}$, a maximum candidate $K_{\max}$.

1. For each candidate $1 \le K \le K_{\max}$,
   - 1.1 (**Label Assignment**) compute the community assignment estimator $\widehat{g}_N$ using spectral clustering on $A^{\mathcal{S}}$;
   - 1.2 (**Parameters Estimation**) calculate the plug-in estimator $\widehat{B}$ defined in (2.6);
   - 1.3 (**SM-BIC Calculation**) calculate the SM-BIC $\widehat{\ell}(K)$, defined in (2.7).
2. Calculate $\widehat{K} = \mathrm{argmax}_{1\le K\le K_{\max}}\widehat{\ell}(K)$.

Output: the optimal choice of the number of communities, $\widehat{K}$.

In the framework of the DCSBM, we need to modify the parameters' estimation methods. First, under candidate DCSBM_K, to obtain $\widehat{g}_N$, we use the spherical spectral clustering method proposed by [49]. Specifically, let v_i be the i-th row of V, i.e., $V = (v_1, \ldots, v_N)^\top$. Furthermore, let $\widetilde{V}$ be the row-normalized version of V, namely, the i-th row of $\widetilde{V}$ is $v_i/\|v_i\|$, where $\|\cdot\|$ denotes the Euclidean norm of a vector. Then, we estimate the node assignments by the following steps: (1) form matrix $\widetilde{V}$ by normalizing each row of V to unit norm; and (2) perform K-means clustering to the rows of $\widetilde{V}$ to obtain $\widehat{g}_N$. Second, based on the embedding matrix V, the plug-in estimator of ψ_i is provided as $\widehat{\psi}_i = \|v_i\|$. Third, for $1 \le k \le l \le K$, the estimated (k,l)-th entry of B is given by,

$$\widehat{B}_{kl} = \frac{o_{kl,\widehat{g}_N}}{n_{kl,\widehat{g}_N}(\widehat{\psi})} = \frac{\sum_{(i,s_j)\in E} A^{\mathcal{S}}_{ij}\mathbb{I}(\widehat{g}_{N,i}=k,\widehat{g}_{N,s_j}=l)}{\sum_{(i,s_j)\in E}\widehat{\psi}_i\widehat{\psi}_{s_j}\mathbb{I}(\widehat{g}_{N,i}=k,\widehat{g}_{N,s_j}=l)}, \tag{2.8}$$

and then take $\widehat{B}_{lk} = \widehat{B}_{kl}$. To this end, we obtain the SM-BIC for DCSBM_K by taking $(\widehat{g}_N, \widehat{B}, \widehat{\psi})$ into (2.5).

Algorithm 2 Model Selection Algorithm for DCSBM.

Input: adjacency matrix $A^{\mathcal{S}}$, a maximum candidate $K_{\max}$.

1. For each candidate $1 \le K \le K_{\max}$,
   - 1.1 (**Label Assignment**) compute the membership labels estimator $\widehat{g}_N$ by performing spherical spectral clustering on $A^{\mathcal{S}}$;
   - 1.2 (**Parameter Estimation**) obtain $\widehat{\psi}$ and $\widehat{B}$ by the following steps,
     - (a) compute the Euclidean norm of each row of matrix V, and then $\widehat{\psi}_i = \|v_i\|$ for all $1 \le i \le N$;
     - (b) calculate the plug-in estimator defined in (2.8);
   - 1.3 (**SM-BIC Calculation**) calculate the SM-BIC $\widehat{\ell}(K)$, defined in (2.5).
2. Calculate $\widehat{K} = \text{argmax}_{1 \le K \le K_{\max}} \widehat{\ell}(K)$.

Output: the optimal choice of the number of communities, $\widehat{K}$.

For convenience, we provide the model selection procedure for the SBM and DCSBM in Algorithms 1 and 2, respectively. To illustrate the model selection algorithm, we show the procedure of identifying K_0 for SBM in Figure 2. Moreover, based on the works of [49, 23], we demonstrated the consistency of spectral clustering for the sub-adjacency matrix $A^{\mathcal{S}}$ in the supplementary materials. To show the effectiveness of the SM-BIC, we discuss its computational complexity in Proposition 1.

Proposition 1 (Computational complexity)**.** *Suppose that the subset of nodes $\mathcal{S}$ is collected by simple random sampling n nodes from $[N]$ without replacement. Then, for both the SBM and DCSBM, the computational complexity of identifying K_0 by SM-BIC is $O(Nn)$.*

The proof of Proposition 1 is provided in Appendix B.2. Note that for each candidate K, in the spectral clustering algorithm, we perform a truncated SVD to the sub-adjacency matrix, where the truncated SVD only computes the largest K eigenvalues and the corresponding eigenvectors with computational complexity $O(Nn)$ for a constant K [28, 55]. Proposition 1 shows the computational advantage of the SM-BIC for large-scale networks. In the next section, we demonstrate that the required subsample size n could be as small as $c(\log N)^2$, where $c > 0$ is a constant. In this case, the computational cost for identifying K_0 based on the SM-BIC could be $O\{N(\log N)^2\}$.

3. Theoretical properties

In this section, we discuss the theoretical properties of the SM-BIC. We first introduce some necessary conditions and subsequently discuss the required subsample size to ensure the effectiveness of the selected sample. Then, we demonstrate the consistency of the SM-BIC under the SBM and DCSBM. Namely, the criterion chooses the right K_0 with probability tending to one as N goes to infinity.

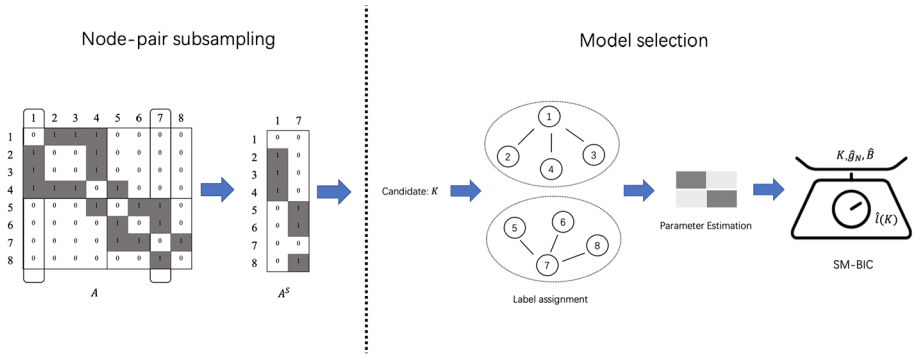

FIG 2. *An illustration of the steps to identify K_0 for SBM based on SM-BIC.*

3.1. Basic assumptions and required subsampling size

To discuss the theoretical properties of the SM-BIC, the following assumptions are considered.

(A1) (Network density) Assume $B^* = \rho_N \widetilde{B}^*$, where $\widetilde{B}^* \in (0,1)^{K_0 \times K_0}$ is a constant matrix and $\rho_N \to 0$ at a rate of $\rho_N N / \log N \to \infty$.

(A2) (Balance level) There exists a constant $c > 0$, such that $\min_{1 \le k \le K_0} N_{k,g_N^*} \ge cN$.

(A3) (Identification condition) Connectivity matrix $B^* \in (0,1)^{K_0 \times K_0}$ has no identical columns.

Assumption (A1) allows for sparse networks, where the network density is $\rho_N \to 0$ at the same rate as in the studies of [78], [41], and [51]. Assumption (A2) requires the size of each community to be relatively balanced. This is a mild and common condition. For example, if the community assignment g_N^* is generated from a multinomial distribution with parameters $\pi = (\pi_1, \ldots, \pi_{K_0})$ such that $\min_{1 \le k \le K_0} \pi_k \ge c/K_0$, then Assumption (A2) is satisfied almost surely. This restriction is also used in [48] and [18]. Assumption (A3) means that the underlying model has K_0 blocks and cannot be further collapsed into a smaller model. It is important to note that the theoretical analysis does not have assortative constraints.

It is noteworthy that a small subsample leads to higher computational efficiency. However, if the subsample size is too small, it is difficult to guarantee the statistical validity of the proposed method. Therefore, we provide two necessary conditions to establish the lower bound of the subsample size n. First, we require that the subsampled nodes cover all blocks with high probability. Specifically, under SBM_{K_0}, we define a set $\mathcal{M}_{K_0} = \{\mathcal{S} : \forall\, k \in [K_0],\ \exists\, i \in \mathcal{S}\ s.t.,\ g_{N,i}^* = k\}$, where $g_{N,i}^*$ is the ground truth label of node i. This implies that the elements in $\mathcal{M}_{K_0}$ completely cover K_0 blocks. Second, we require that the average degree of the subnetwork should increase with N. Specifically, let $d_i = \sum_{1 \le j \le n} A_{ij}^S$ denote the degree of node i in the subnetwork based on $A^{\mathcal{S}}$, for $i = 1, \ldots, N$.

Furthermore, let $d = \sum_{i=1}^{N} d_i/N$ denote the average degree of the subnetwork. Then, we assume the expected average degree $\mathbb{E}(d) = \Omega(\log N)$. Based on these two conditions, we provide the lower bound of subsample size n in the following proposition.

Proposition 2 (Subsample size)**.** *Under Assumptions (A1)–(A3), suppose $\mathcal{S}$ is collected by simple random sampling n nodes from the entire network without replacement. If the subsample size is $n = \Omega(\log N/\rho_N)$, then we have $\mathcal{S} \in \mathcal{M}_{K_0}$ and $\mathbb{E}(d) = \Omega(\log N)$ with probability at least $1 - 1/N$.*

The proof is provided in Appendix B.3. Note that $n = \Omega(\log N/\rho_N)$ means that there are positive constants c and N_0 such that $n \geq c \log N/\rho_N$ for all $N > N_0$ [46]. According to Proposition 2, the lower bound of subsample size goes to infinity with a lower speed compared to N. In particular, consider $\rho_N = (\log N)^{-1}$, then the subsample size $n = \Omega\{(\log N)^2\}$.

Based on Proposition 2, we discuss the proportion of utilized edges in the proposed SM-BIC method. Given that the subsampling size is $n = \Omega(\log N/\rho_N)$, the number of utilized edges in the SM-BIC algorithm is on the order of $N \log N$. Additionally, the total number of connections available in the entire network is on the order of $N^2\rho_N$. Consequently, the proportion of utilized edges is $\log N/(N\rho_N)$, which can be further expressed as $(\log N)^2/N$ for $\rho_N = (\log N)^{-1}$. In the realm of large-scale networks, SM-BIC effectively reduces computational costs, albeit at the expense of losing some observed samples.

We then explore the theoretical properties of the proposed criteria. Before the theoretical analysis, we introduce two key remarks. Remark 1 discusses the gap between the proposed criteria and the estimated criteria. Meanwhile, Remark 2 demonstrates the challenges of considering the randomness introduced by the sampling step in theoretical analysis.

Remark 1 (Likelihood Estimation in Algorithms)**.** *We propose Algorithms 1 and 2 to estimate the likelihood of SBM and DCSBM, respectively. The primary divergence between the maximum likelihood and estimated likelihood arises from label vector estimation. Notably, the consistency of spectral clustering for subsampled networks has been established by [18, 24], and the details of the proof are also provided in the supplementary material. In this way, we proceed to analyze the consistency of SM-BIC based on the maximum likelihood.*

Remark 2 (Randomness Introduced by the Sampling Step)**.** *The subsampled adjacency matrix is obtained through random sampling from the entire adjacency matrix. We gauge the additional randomness by comparing the log-likelihood of the subsampled adjacency matrix to the regularized log-likelihood of the entire adjacency matrix. It can be represented as* $\max_{g_N \in \mathcal{C}(A^{\mathcal{S}}, K)} \sup_{\theta \in \Theta_K} \log f(A^{\mathcal{S}}|g_N, \theta) - \frac{2M}{N(N-1)} \max_{g_N \in [K]^N} \sup_{\theta \in \Theta_K} \log f(A|g_N, \theta)$*. To theoretically control this divergence, stricter conditions for the network model are required. However, as the sample size n increases, the expectation of this divergence approaches zero and shows lower variance. Therefore, in the following discussion, we focus on the subsampled adjacency matrix $A^{\mathcal{S}}$ with its sample size satisfying the condition in Proposition 2.*

3.2. Consistency of SM-BIC

We first establish the consistency of the SM-BIC under SBMs. Given a subsampled adjacency matrix $A^{\mathcal{S}}$, the underlying SM-BIC of $\mathrm{SBM}_{K_0}(g_N^*, B^*)$ is

$$\ell^*(K_0) = \log f(A^{\mathcal{S}}|g_N^*, B^*) - \left\{ n \log K_0 + \frac{K_0(K_0+1)}{4} \log M \right\}.$$

Intuitively, fitting the observed network with a correct number of communities yields the largest value of the SM-BIC. Then, for any candidate SBM_K, we compare its SM-BIC $\ell(K)$ with the underlying SM-BIC $\ell^*(K_0)$ under three different cases, namely, under-fitting ($K < K_0$), correctly fitting ($K = K_0$), and over-fitting ($K > K_0$). We analyze the divergence between $\ell(K)$ and $\ell^*(K_0)$, which is

$$\begin{aligned} \ell(K) - \ell^*(K_0) &= \left\{ \max_{g_N \in \mathcal{C}(A^{\mathcal{S}}, K)} \sup_{B \in \mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) - \log f(A^{\mathcal{S}}|g_N^*, B^*) \right\} \\ &\quad - \left\{ n \log (K/K_0) + \frac{K(K+1) - K_0(K_0+1)}{4} \log M \right\} \\ &= L_{K,K_0} - R_{K,K_0}, \end{aligned} \tag{3.1}$$

where $L_{K,K_0} = \max_{g_N \in \mathcal{C}(A^{\mathcal{S}}, K)} \sup_{B \in \mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) - \log f(A^{\mathcal{S}}|g_N^*, B^*)$, and $R_{K,K_0} = n \log (K/K_0) + \{K(K+1) - K_0(K_0+1)\}/4 \log M$. It is noteworthy that L_{K,K_0} is a log-likelihood ratio, which measures the goodness-of-fit of the estimated model compared with the underlying model. Since R_{K,K_0} is fixed for a given K and n, we focus on analyzing L_{K,K_0} in the three cases mentioned above.

Case 1: Under-fitting. In this case, we prove the upper bound for the log-likelihood ratio L_{K,K_0} in the following theorem.

Theorem 1 (Upper bound of the log-likelihood ratio under under-fitting)**.** *Suppose A is generated from $\mathrm{SBM}_{K_0}(g_N^*, B^*)$. Furthermore, suppose Assumptions (A1)–(A3) hold and n satisfies the condition in Proposition 2. If $K < K_0$, then $L_{K,K_0} = -\Omega_P(\rho_N M)$.*

The technical proof of Theorem 1 can be found in Appendix C.1. For $K < K_0$, it can be verified that $R_{K,K_0} = -\Omega(n + \log M)$. Combining the conclusion in Theorem 1, we have $\ell(K) - \ell^*(K_0) = -\Omega_P(\rho_N M)$ by (3.1). Moreover, note that the lower bound of the ratio L_{K,K_0} is negatively related to ρ_N and M, and goes to negative infinity as $N \to \infty$. This indicates that under the proposed conditions, the SM-BIC avoids the under-fitting case with high probability.

Case 2: Correctly fitting. We then analyze the log-likelihood ratio L_{K,K_0} under a given correct number of communities, i.e., $K = K_0$.

Theorem 2 (Convergence of the log-likelihood ratio under SBM)**.** *Make the same assumptions as in Theorem 1. If $K = K_0$, then we have $L_{K_0,K_0} = O_P(\rho_N)$.*

The proof is provided in Appendix C.2. When $K = K_0$, since $R_{K_0,K_0} = 0$, together with the conclusion in Theorem 2, we have $\ell(K) - \ell^*(K_0) = L_{K_0,K_0} = O_P(\rho_N)$. Moreover, in Case 2, Theorem 2 implies that the log-likelihood ratio L_{K_0,K_0} converges faster in sparse networks.

Case 3: Over-fitting. Similar to the conclusion in Case 1, we present the upper bound of the log-likelihood ratio L_{K,K_0} in the following theorem.

Theorem 3 (Upper bound of log-likelihood ratio under over-fitting)**.** *Make the same assumptions as in Theorem 1. For any candidate $K > K_0$, we have $L_{K,K_0} = O_P(\log N)$.*

The proof of this theorem can be found in Appendix C.3. For $K > K_0$, by the definition of R_{K,K_0}, we have $R_{K,K_0} = \Omega(n + \log M)$. Then, together with the conclusion in Theorem 3, we have $\ell(K) - \ell^*(K_0) = -\Omega_P(n + \log M)$. Note that the upper bound is negatively related to the subsample size n, which indicates that the SM-BIC avoids over-fitting with increasing probability as n grows. Therefore, Theorem 3 ensures that the subsample size $n = \Omega(\log N/\rho_N)$ is large enough to prevent this misspecification.

To summarize, we establish the consistency of the SM-BIC under the SBM in the following corollary.

Corollary 1 (Consistent results for SBM)**.** *Suppose A is generated from $\mathrm{SBM}_{K_0}(g_N^*, B^*)$ and Assumptions (A1) and (A2) hold. If the subsample size n satisfies the condition in Proposition 2, then for $K \neq K_0$, we have $P(\ell(K) > \ell^*(K_0)) \to 0$, with $N \to \infty$.*

Corollary 1 demonstrates that for the SBM, the correct number of communities can be identified by the SM-BIC with high probability.

Now, we investigate the consistency of the SM-BIC under the DCSBM. We assume that the degree heterogeneity parameter ψ is known, which is also considered in the theoretical studies of [48] and [30]. In this case, according to criterion (2.5), we have $L_{K,K_0} = \max_{g_N \in \mathcal{C}(A^{\mathcal{S}},K)} \sup_{B \in \mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B, \psi^*) - \log f(A^{\mathcal{S}}|g_N^*, B^*, \psi^*)$. Then, we first investigate the convergence of the log-likelihood ratio under the correct specification.

Theorem 4 (Convergence of the log-likelihood ratio under DCSBM)**.** *Suppose that A is generated from $\mathrm{DCSBM}_{K_0}(g_N^*, B^*, \psi^*)$. Under Assumptions (A1) and (A2), if n satisfies the condition in Proposition 2, for $K = K_0$, we have $L_{K,K_0} = O_P(\rho_N)$.*

The proof of Theorem 4 is provided in Appendix C.4. According to Theorem 4, under the DCSBM, the convergence of L_{K,K_0} can also be guaranteed if K is correctly specified.

Based on Theorem 4, together with similar arguments, one can show that the conclusions of Theorem 1 and Theorem 3 hold under the DCSBM. Hence, we draw the theoretical results for the DCSBM as follows.

Corollary 2 (Consistent results for DCSBM)**.** *Suppose A is generated from $\mathrm{DCSBM}_{K_0}(g_N^*, B^*, \psi^*)$ and Assumptions (A1)–(A3) hold. If n satisfies the con-*

dition in Proposition 2, for $K \neq K_0$, we have $P(\ell(K) > \ell^*(K_0)) \to 0$, with $N \to \infty$.

4. Numerical studies

4.1. Simulation models and performance measurements

We start with the generation mechanism of the networks. For a given K_0, we assume that the underlying node labels are generated by $g^*_{N,i} \sim \text{Multinomial}(\pi)$ independently for all $i = 1, \ldots, N$, where $\pi = (1/K_0, \ldots, 1/K_0)$. Second, we define the connectivity matrix as $B^* = \rho_N(\beta \mathbf{1}_{K_0}\mathbf{1}_{K_0}^\top + (1-\beta)I_{K_0})$, where $\mathbf{1}_{K_0} \in \mathbb{R}^{K_0}$ is filled with elements 1 and $I_{K_0} \in \mathbb{R}^{K_0 \times K_0}$ is an identity matrix, and the *out-in-ratio* parameter $\beta \in (0, 1)$ measures the connectivity divergence within and between communities.

Then, we evaluate the performance of the SM-BIC through the following three different examples under SBM framework.

Example 1 (Consistency of the approximated SM-BIC)**.** *Let the number of communities K_0 vary from 2 to 5. For each K_0, let N increase from 500 to 5,000. Furthermore, set $\beta = 0.15$, $\rho_N = N^{-1/2}$, $n = \lceil \zeta \log N/\rho_N \rceil$, where $\lceil x \rceil$ represents the smallest integer of no less than x. The parameter ζ is set to 1.0, 1.5, and 2.0, respectively. Additionally, we examine the difference between $\ell(K_0)$ and $\hat{\ell}(K_0)$, with sample size n increasing while keeping $N = 5{,}000$ fixed.*

Example 2 (The effect of network density)**.** *Let the number of communities K_0 vary from 2 to 5 and the entire network size N increase from 1,000 to 5,000. Additionally, take the out-in-ratio parameter as $\beta = 0.15$ and let the network density ρ_N increase from $0.5N^{-1/2}$ to $1.5N^{-1/2}$. For each network setting, we take the subsample size as $n = \lceil 1.5N^{1/2} \log N \rceil$.*

Example 3 (The effect of arbitrary outlier nodes)**.** *According to the generalized stochastic block model proposed by [13], we generate networks with a portion of outlier nodes. Specifically, assume that there are N normal nodes and m outlier nodes. The connections between N normal nodes obey the $\text{SBM}_{K_0}(g^*_N, B^*)$ with $\beta = 0.15$ and $\rho_N = N^{-1/2}$, while connections between outliers are generated from a random graph model with a connectivity probability of 0.1. Moreover, define X as a $N \times m$ matrix with independent Bernoulli entries, representing the connections between normal nodes and outlier nodes. Let $\mathbb{E}X = \boldsymbol{v}\mathbf{1}_m^\top$ where the components of $\boldsymbol{v}$ are N i.i.d. copies of $\boldsymbol{u}^2/10$ and $\boldsymbol{u}$ is a uniform random variable on $[0, 1]$. Furthermore, let m increase from 20 to 100.*

To further evaluate the performance of the SM-BIC method, we compare it with four existing approaches, namely, the method based on the Bethe-Hessian matrix with moment correction (BHMC) proposed by [47], the network cross-validation (NCV) method proposed by [18], the network cross-validation method by edge sampling (ECV) proposed by [51], and the corrected Bayesian information criterion (CBIC) proposed by [41].

Example 4 (Comparison under SBM)**.** *We generated the network from* $\mathrm{SBM}_{K_0}(g_N^*, B^*)$ *with* $\beta = 0.35$ *and* $\rho_N = 20\log N/N$*. Furthermore, let the network size* N *increase from 3,000 to 10,000 and* K_0 *vary from 2 to 5, accordingly.*

Example 5 (Comparison under DCSBM)**.** *We follow the scenario proposed in [86]. The parameters* ψ_i *are independently generated from a distribution with expectation 1, specifically,*

$$\psi_i = \begin{cases} \eta_i, & \text{with probability } \alpha; \\ 1/3, & \text{with probability } (1-\alpha)/2; \\ 5/3, & \text{with probability } (1-\alpha)/2, \end{cases}$$

where η_i *is uniformly distributed on the interval* $[3/5, 7/5]$*. The variance of* ψ_i *is equal to* $4\alpha/75 + 4(1-\alpha)/9$*. Then, the variance is a decreasing function of* α*. We vary* α *from 0.4 to 0.8. The other parameters are set to be the same in Example 4.*

Throughout this simulation study, we set the maximum candidate to $K_{\max} = 10$. The random experiments are repeated $T = 100$ times to ensure a reliable evaluation. Additionally, for each repetition, we assume the selected number of communities is $\widehat{K}_t$, for $t = 1, \ldots, T$. Then, to gauge the performance of the SM-BIC, we consider two measurements. First, the probability of correct identification is defined as

$$\text{Prob} = \sum_{t=1}^{T} \mathbb{I}(\widehat{K}_t = K_0)/T, \tag{4.1}$$

where a larger Prob corresponds to more accurate model selection. Second, the average of the selected number of communities is defined by

$$\text{Mean} = \sum_{t=1}^{T} \widehat{K}_t/T. \tag{4.2}$$

All simulations are conducted in a Linux server with a 3.60 GHz Intel Core i7-9700K CPU and 16 GB RAM.

4.2. Simulation results

All simulation results are shown in Tables 1–5 and Figure 4. We draw the following conclusions from different examples.

Example 1. The simulation results are presented in Table 1 and Figure 3. We make the following comments. First, as n grows from $\lceil \log N/\rho_N \rceil$ to $\lceil 2\log N/\rho_N \rceil$, the probability of correct identification increases from 0.84 to 1.00 under the setting $K_0 = 5$ and $N = 500$. Second, as N increases from 500 to 5,000, the probability of correct identification increases from 0.84 to 1.00 under the setting $K_0 = 5$ and $n = \lceil \log N/\rho_N \rceil$. Third, as the network size N increases from 500

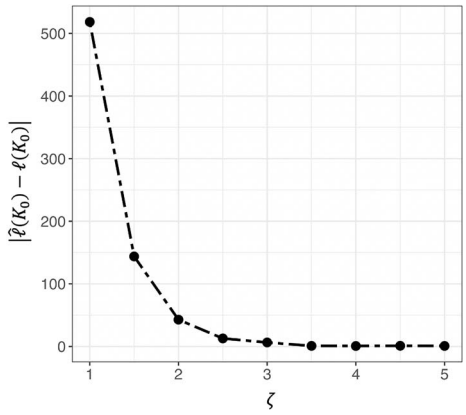

FIG 3. *The difference between $\hat{\ell}(K_0)$ and $\ell(K_0)$ decreases as ζ increases, with the sample size $n = \lceil \zeta \log N/\rho_N \rceil$ growing, while maintaining a fixed network size of $N = 5{,}000$ and a true number of communities of $K_0 = 2$.*

TABLE 1

Simulation results for Example 1 with network density $\rho_N = N^{-1/2}$ and subsample size $n = \lceil \zeta \log N/\rho_N \rceil$. Measurements include "Prob" and "Mean" (defined in (4.1) and (4.2)). Average CPU computational time (in seconds) is also provided.

		$N = 500$			$N = 2{,}000$			$N = 5{,}000$		
K_0	ζ	Prob	Mean	CPU	Prob	Mean	CPU	Prob	Mean	CPU
2	1.0	1.00	2.00	0.73	1.00	2.00	2.08	1.00	2.00	6.77
	1.5	1.00	2.00	0.76	1.00	2.00	2.41	1.00	2.00	8.51
	2.0	1.00	2.00	0.78	1.00	2.00	2.71	1.00	2.00	10.70
3	1.0	1.00	3.00	0.69	1.00	3.00	1.93	1.00	3.00	6.36
	1.5	1.00	3.00	0.71	1.00	3.00	2.19	1.00	3.00	7.97
	2.0	1.00	3.00	0.72	1.00	3.00	2.45	1.00	3.00	9.93
4	1.0	0.96	4.00	0.65	1.00	4.00	1.82	1.00	4.00	6.14
	1.5	0.99	4.01	0.67	1.00	4.00	2.03	1.00	4.00	7.67
	2.0	1.00	4.00	0.67	1.00	4.00	2.26	1.00	4.00	9.55
5	1.0	0.84	4.86	0.62	1.00	5.00	1.73	1.00	5.00	5.99
	1.5	1.00	5.00	0.64	1.00	5.00	1.95	1.00	5.00	7.49
	2.0	1.00	5.00	0.65	1.00	5.00	2.16	1.00	5.00	9.20

to 5,000, the average CPU computational time of each experiment does not exceed 10.70 seconds. Hence, the SM-BIC is an efficient and consistent method for large-scale networks, and these results are consistent with our theoretical results in Proposition 2 and Corollary 1. Moreover, Figure 3 illustrates the convergence of the difference $|\hat{\ell}(K_0) - \ell(K_0)|$ as the sample size n increases.

EXAMPLE 2. The simulation results are provided in Table 2. We obtain the following findings. First, as network density ρ_N increases from $0.5N^{-1/2}$ to $1.5N^{-1/2}$, the probability of correct identification increases to 1 for all $K_0 = 2, \ldots, 5$. Second, even in the sparsest case $\rho_N = 0.5N^{-1/2}$, as N grows from 1,000 to 5,000, the probability of correct identification increases from 0.75 to 1.00. Hence, for large-scale networks, the proposed method allows for a higher level of sparsity.

TABLE 2

Simulation results for Example 2, with network subsample size $n = \lceil 1.5N^{1/2}\log N\rceil$. Measurements include "Prob" and "Mean" (defined in (4.1) and (4.2)). Average CPU computational time (in seconds) is also reported.

		$N = 1{,}000$			$N = 3{,}000$			$N = 5{,}000$		
$\rho_N N^{1/2}$	K_0	Prob	Mean	CPU	Prob	Mean	CPU	Prob	Mean	CPU
0.5	2	1.00	2.00	1.24	1.00	2.00	3.74	1.00	2.00	8.41
	3	1.00	3.00	1.16	1.00	3.00	3.56	1.00	3.00	8.02
	4	0.99	3.99	1.13	1.00	4.00	3.41	1.00	4.00	7.99
	5	0.75	4.75	1.11	1.00	5.00	3.28	1.00	5.00	7.73
1.0	2	1.00	2.00	1.26	1.00	2.00	3.88	1.00	2.00	8.63
	3	1.00	3.00	1.15	1.00	3.00	3.53	1.00	3.00	8.04
	4	1.00	4.00	1.08	1.00	4.00	3.31	1.00	4.00	7.71
	5	1.00	5.00	1.03	1.00	5.00	3.19	1.00	5.00	7.59
1.5	2	1.00	2.00	1.21	1.00	2.00	3.92	1.00	2.00	8.85
	3	1.00	3.00	1.08	1.00	3.00	3.48	1.00	3.00	7.96
	4	1.00	4.00	1.00	1.00	4.00	3.25	1.00	4.00	7.56
	5	1.00	5.00	0.95	1.00	5.00	3.07	1.00	5.00	7.33

TABLE 3

Simulation results for Example 3, with network density $\rho_N = N^{-1/2}$ and subsample size $n = \lceil 1.5\log N/\rho_N\rceil$. Outlier nodes ($m$) range from 20 to 100 for each network with N nodes. Measurements include "Prob" and "Mean" (defined in (4.1) and (4.2)). Average CPU computational time (in seconds) is also reported.

		$N = 2{,}000$			$N = 3{,}000$			$N = 5{,}000$		
m	K_0	Prob	Mean	CPU	Prob	Mean	CPU	Prob	Mean	CPU
20	2	1.00	2.00	2.36	1.00	2.00	3.79	1.00	2.00	7.80
	3	1.00	3.00	2.15	1.00	3.00	3.52	1.00	3.00	7.34
	4	1.00	4.00	2.00	1.00	4.00	3.15	1.00	4.00	7.26
	5	1.00	5.00	1.89	1.00	5.00	3.00	1.00	5.00	7.02
50	2	1.00	2.00	2.39	1.00	2.00	3.71	1.00	2.00	7.59
	3	1.00	3.00	2.17	1.00	3.00	3.44	1.00	3.00	7.41
	4	1.00	4.00	2.00	1.00	4.00	3.06	1.00	4.00	7.24
	5	1.00	5.00	1.87	1.00	5.00	2.89	1.00	5.00	6.88
100	2	1.00	2.00	2.42	1.00	2.00	3.59	1.00	2.00	7.79
	3	1.00	3.00	2.19	1.00	3.00	3.41	1.00	3.00	7.62
	4	0.97	4.03	2.02	1.00	4.00	3.20	1.00	4.00	7.30
	5	0.82	5.18	1.89	0.99	5.01	2.95	1.00	5.00	6.90

EXAMPLE 3. This simulation results are provided in Table 3. We draw the following conclusions. First, as the number of outliers decreases from 100 to 20, the accuracy of recovering K_0 increases from 0.82 to 1.00 under the setting $N =$ 2,000 and $K_0 = 5$. Second, as N varies from 2,000 to 5,000, the probability of correct identification grows from 0.82 to 1.00 in the case of $K_0 = 5$. Therefore, for large-scale networks with arbitrary outliers, the SM-BIC method can accurately identify the number of communities with high probability.

EXAMPLE 4. The comparison results are shown in Table 4 and Figure 4. We draw the following conclusions. First, SM-BIC is more accurate than the ECV method in this study. Specifically, for the setting of $N = 3{,}000$, when $K_0 = 4$ and $K_0 = 5$, the Prob of the ECV method is only 0.87 and 0.80, respectively, while the Prob of the SM-BIC is 1.00 in these cases. Second, the average computational

TABLE 4

Simulation results for Example 4, with network density $\rho_N = 20 \log N/N$ and subsample size $n = \lceil 1.5 \log N/\rho_N \rceil$. Measurements include "Prob" and "Mean" (defined in (4.1) and (4.2)). Average CPU computational time (in seconds) is also reported.

		$N = 3{,}000$			$N = 5{,}000$			$N = 10{,}000$		
K_0	Method	Prob	Mean	CPU	Prob	Mean	CPU	Prob	Mean	CPU
2	BHMC	1.00	2.00	21.40	1.00	2.00	105.51	1.00	2.00	463.31
	NCV	1.00	2.00	52.88	1.00	2.00	190.88	1.00	2.00	533.37
	ECV	1.00	2.00	220.76	1.00	2.00	685.66	1.00	2.00	2064.47
	CBIC	1.00	2.00	12.92	0.97	2.07	36.42	1.00	2.00	70.51
	SMBIC	1.00	2.00	1.66	1.00	2.00	3.25	1.00	2.00	12.43
3	BHMC	1.00	3.00	21.80	1.00	3.00	107.49	1.00	3.00	524.90
	NCV	1.00	3.00	45.34	1.00	3.00	164.78	1.00	3.00	512.04
	ECV	0.97	3.03	208.78	0.97	3.03	654.23	1.00	3.00	2112.26
	CBIC	1.00	3.00	12.59	1.00	3.00	34.72	1.00	3.00	71.36
	SMBIC	1.00	3.00	1.52	1.00	3.00	3.14	1.00	3.00	10.81
4	BHMC	1.00	4.00	21.79	1.00	4.00	107.42	1.00	4.00	525.26
	NCV	1.00	4.00	40.65	0.97	4.07	149.09	1.00	4.00	461.67
	ECV	0.87	4.30	203.66	0.93	4.10	621.99	1.00	4.00	2035.38
	CBIC	1.00	4.00	12.27	1.00	4.00	34.10	1.00	4.00	74.47
	SMBIC	1.00	4.00	1.43	1.00	4.00	2.98	1.00	4.00	10.05
5	BHMC	1.00	5.00	21.81	1.00	5.00	107.82	1.00	5.00	525.84
	NCV	1.00	5.00	37.03	0.97	5.07	134.99	1.00	5.00	423.69
	ECV	0.80	5.30	202.86	0.80	5.27	639.31	0.80	5.40	2033.33
	CBIC	1.00	5.00	12.35	1.00	5.00	34.11	1.00	5.00	75.97
	SMBIC	1.00	5.00	1.35	1.00	5.00	2.74	1.00	5.00	9.50

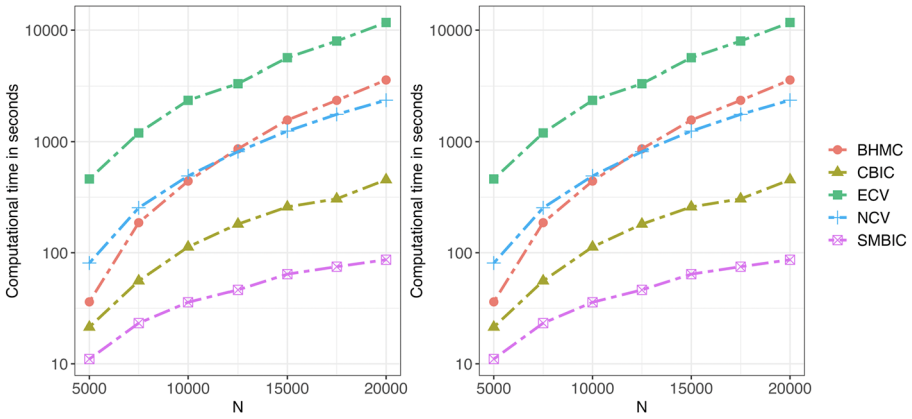

FIG 4. *The computational time (in seconds) for model selection methods is examined under the settings outlined in Example 4. The y-axis is presented in a logarithmic scale for clarity. We report the average CPU time for two simulation scenarios: when the number of communities is $K_0 = 2$ (left panel) and when $K_0 = 5$ (right panel).*

time of the SM-BIC is much smaller than that of the BHMC, NCV, ECV, and CBIC, especially when N is large. As shown in Figure 4, the average CPU computational time of these methods is further compared across diverse network sizes. We observe that the average CPU computational time of the SM-BIC is the

TABLE 5

Simulation results for Example 5, with network density $\rho_N = 20 \log N/N$ and subsample size $n = \lceil 1.5 \log N/\rho_N \rceil$. The heterogeneity parameter α varies from 0.4 to 0.8. Measurements include "Prob" and "Mean" (defined in (4.1) and (4.2)). Average CPU computational time (in seconds) is also reported.

		$\alpha = 0.4$			$\alpha = 0.6$			$\alpha = 0.8$		
K_0	Method	Prob	Mean	CPU	Prob	Mean	CPU	Prob	Mean	CPU
2	BHMC	1.00	2.00	51.10	1.00	2.00	50.81	1.00	2.00	51.16
	NCV	1.00	2.00	120.01	1.00	2.00	119.78	1.00	2.00	118.66
	ECV	0.00	3.77	545.87	0.00	5.30	545.81	0.23	7.23	549.83
	CBIC	1.00	2.00	38.46	1.00	2.00	38.58	0.93	2.13	38.57
	SMBIC	1.00	2.00	16.63	1.00	2.00	16.51	1.00	2.00	16.39
3	BHMC	1.00	3.00	50.34	1.00	3.00	51.01	1.00	3.00	51.75
	NCV	1.00	3.00	108.66	1.00	3.00	108.72	1.00	3.00	107.80
	ECV	0.00	4.00	534.00	0.00	4.00	540.07	0.00	7.93	542.12
	CBIC	1.00	3.00	37.37	1.00	3.00	37.02	1.00	3.00	36.88
	SMBIC	1.00	3.00	16.44	1.00	3.00	16.41	1.00	3.00	16.29
4	BHMC	1.00	4.00	50.40	1.00	4.00	51.28	1.00	4.00	52.09
	NCV	1.00	4.00	101.31	1.00	4.00	101.83	1.00	4.00	100.58
	ECV	0.00	5.10	526.65	0.00	5.17	532.16	0.00	5.23	532.51
	CBIC	1.00	4.00	36.11	1.00	4.00	36.27	1.00	4.00	36.23
	SMBIC	1.00	4.00	16.27	1.00	4.00	16.29	1.00	4.00	16.12
5	BHMC	1.00	5.00	49.92	1.00	5.00	51.50	1.00	5.00	52.54
	NCV	0.97	5.03	96.69	1.00	5.00	97.35	0.97	5.03	95.12
	ECV	0.00	6.67	529.87	0.00	7.27	532.77	0.00	6.87	534.82
	CBIC	1.00	5.00	36.11	1.00	5.00	36.46	1.00	5.00	36.10
	SMBIC	1.00	5.00	16.13	1.00	5.00	16.09	1.00	5.00	16.00

smallest, while the ECV method is much more computationally expensive than other algorithms. Because in each iteration, ECV performs matrix completion and estimates community labels from a $N \times N$-dimensional low-rank matrix.

EXAMPLE 5. The comparison results are reported in Table 5. We draw the following conclusions. First, the SM-BIC method can correctly identify K_0 with α varying from 0.4 to 0.8, while the ECV method shows lower accuracy than other approaches. Second, compared with the BHMC, NCV, ECV, and CBIC methods, when $\alpha = 0.4$ and $K_0 = 2$, the average CPU computational time for the BHMC, NCV, ECV, and CBIC methods is 51.10s, 120.01s, 545.87s, and 38.46s, respectively, while that of the SM-BIC method is only 16.63s. Thus, in the DCSBM, the SM-BIC is more robust than the ECV method in terms of degree heterogeneity and more computationally efficient than all these methods.

4.3. Real data analysis

Political blog dataset The political blog dataset was collected and analyzed in [1]. The data set consists of over one thousand blogs discussing US politics, with edges representing web links. The nodes are labeled as being either "conservative" or "liberal", which can be treated as two well-defined communities. We only consider the largest connected component of this network, which consists of 1,222 nodes with community sizes of 586 and 636, while the network density is $\rho_N = 2.24\%$. The degree-corrected stochastic block model is

TABLE 6
Comparison results of different methods in the dataset of housing prices in Beijing. The estimated number of communities $\widehat{K}$ and the CPU computational time of each method are reported.

Model		BH	NCV	ECV	CBIC	SMBIC
SBM	$\widehat{K}$	3.00	3.00	3.00	4.00	3.00
	CPU	79.21	70.50	646.21	52.77	8.75
DCSBM	$\widehat{K}$	3.00	3.00	3.00	3.00	3.00
	CPU	77.26	111.76	686.25	132.45	12.58

believed to fit better for this network than stochastic block model [45, 85]. Then, under the DCSBM framework, we take the subsample size of the SM-BIC as $n = \lceil 1.5 \log N/\rho_N \rceil = 475$, and compare the SM-BIC method with other algorithms. Specifically, we obtain the estimated number of communities as 2 by the NCV, CBIC, and SM-BIC, with computation times of 5.57s, 3.82s, and 1.60s, respectively. While the BHMC and ECV estimate $\widehat{K} = 7$ and $\widehat{K} = 6$, respectively. We see that the NCV, CBIC, and SM-BIC methods all give correct estimates for the number of communities, and SM-BIC further outperforms these two algorithms in terms of computational efficiency.

A House price dataset This dataset is publicly available on the platform *Kaggle* (`https://Kaggle.com`), which contains housing transaction information in Beijing from 2011 to 2017. Here, we collect 6,000 samples traded in 2016, distributed in the "Feng Tai", "Chang Ping", and "Hai Dian" districts of Beijing. The nodes are these collected samples and a network is obtained by randomly connecting the node pairs in the same district with a probability of 0.1. That is, if node i and j are in the same district, then we add an edge to node pair (i, j) with probability 0.1. As a result, this network has three well-defined communities with the sizes of communities 1,661, 2,365, and 1,974, respectively, while the network density is $\rho_N = 3.40\%$. We then apply the SM-BIC and the aforementioned methods to identify the number of communities for this network under the SBM and DCSBM frameworks, respectively. For the SM-BIC method, the subsample size is set to be $n = \lceil 2 \log N/\rho_N \rceil = 511$. The results are provided in Table 6. As shown in Table 6, we observe that the SM-BIC method can correctly identify the number of communities under both the SBM and DCSBM frameworks. Moreover, the SM-BIC takes only 8.75s for SBM, which is only 11.0% of BH, 12.4% of NCV, and 1.4% of ECV, respectively. For the DCSBM model, the SM-BIC takes 12.58s, which is only 16.3% of BH, 11.3% of NCV, 1.8% of ECV, and 9.5% of CBIC, respectively.

5. Concluding remarks

This work proposes a subsampling-based modified Bayesian information criterion (SM-BIC) to identify the number of communities for large-scale SBMs. We also extend this criterion to DCSBMs. Specifically, the technical conditions

of subsampling size are derived, and the consistency properties of SM-BIC are established. In the context of large-scale networks, the proposed SM-BIC has more valuable computational advantages than existing model selection methods. Specifically, the computational complexity of the SM-BIC for both the SBM and DCSBM could be as low as $O\{N(\log N)^2\}$. Consequently, the SM-BIC method could be performed even using a personal computer. Numerical studies further demonstrate these computational improvements.

To conclude this work, we consider several interesting topics for future research. First, in this study, we focus on reducing computational costs by network subsampling only once; this idea can be extended to a resampling approach, which is currently under investigation. Second, informative subsamples are important for extracting useful information from the entire network. Subsampling strategies for independent big data have been extensively studied; see [63], [74], and [82] for further discussions. Based on these studies, it would be interesting to investigate subnetwork extraction methods with meaningful statistical interpretations in large-scale networks. Third, in this work, following [54], we assume that K_0 is fixed. However, it is an interesting and challenging question to allow for a diverging K_0. We will work in this direction in future research.

The code is publicly available on GitHub (https://github.com/Stamath/SMBIC).

Appendix A: Necessary notations and lemmas

In Appendix A, we introduce some necessary notations in Appendix A.1. Then, we give three useful lemmas for the subsequent theoretical proof of the proposed method in Appendix A.2.

A.1. Notations

Given a label vector g_N, we define some necessary count statistics. Define a $K \times K$ count matrix as $n_{g_N} = (n_{kl,g_N})_{1\le k,l\le K}$ and $o_{g_N} = (o_{kl,g_N})_{1\le k,l\le K}$. Let $\boldsymbol{p} = (N_{1,g_N^*},\ldots,N_{K,g_N^*})^\top/N$ denote the underlying block proportions, where $N_{k,g_N^*} = \sum_{i=1}^N \mathbb{I}(g_{N,i}^* = k)$ represents the number of nodes belonging to the k-th cluster. For two sets of labels g_N and g'_N, define $|g_N - g'_N| = \sum_{i=1}^N \mathbb{I}(g_{N,i} \ne g'_{N,i})|$. In addition, define τ as a permutation on $[K]$ and denote $\|\cdot\|_\infty$ as a maximum norm of a matrix.

For simplicity, we quote the notations from [78] to characterize the log-likelihood function. Let H_{g_N} be an $K \times K_0$ confusion matrix whose (k,l)-entry is $H_{kl,g_N} = 1/N \sum_{i=1}^N \mathbb{I}\{g_{N,i} = k, g_{N,i}^* = l\}$. Additionally, we define

$$F(Q,q) = \sum_{1\le k\le l\le K} q_{kl}\gamma\left(\frac{Q_{kl}}{q_{kl}}\right),$$

where $\gamma(x) = x\log x + (1-x)\log(1-x)$ for $x \in (0,1)$. Then, for a fixed label vector g_N, the corresponding log-likelihood can be expressed as

$\sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) = MF(o_{g_N}/M, n_{g_N}/M)$. We further define its expectation as

$$G(H_{g_N}, B^*) = \sum_{1\le k\le l\le K} (H_{g_N}\mathbf{1}\mathbf{1}^\top H_{g_N}^\top)_{kl}\gamma\Big\{\frac{(H_{g_N}B^*H_{g_N}^\top)_{kl}}{(H_{g_N}\mathbf{1}\mathbf{1}^\top H_{g_N}^\top)_{kl}}\Big\}.$$

A.2. Useful lemmas

Here, we provide some useful lemmas, that is, Lemmas 2–4, for the proof of the consistency of the SM-BIC.

In statistics, Hoeffding inequality provides an upper bound for the sum of bounded random variables, which was proved by [38].

Lemma 2 (Hoeffding inequality)**.** *Let x_i, $i = 1, \ldots, N$, be mutually independent random variables such that $a_i \le x_i \le b_i$ almost surely. Consider the sum of these random variables, $Y_N = \sum_{i=1}^N x_i$. Then, for all $s > 0$,*

$$P\{Y_N - E(Y_N) \ge s\} \le \exp\Big\{-\frac{2s^2}{\sum_{i=1}^N (b_i - a_i)^2}\Big\}.$$

In the under-fitting case, without loss of generality, we start with $K = K_0 - 1$, and the following Lemma 3 shows that $G(H_{g_N}, B^*)$ is maximized by combining two existing communities in g_N^*.

Lemma 3 (Expectation of the log-likelihood function of under-fitting)**.** *Given the true label g_N^*, suppose $g_N \in \mathcal{C}(A^{\mathcal{S}}, K_0 - 1)$, and then maximizing the function $G(H_{g_N}, B^*)$ over H_{g_N} achieves its maximum in the label set*

$$\{g_N \in \mathcal{C}(A^{\mathcal{S}}, K_0 - 1) : \text{ there exists } \tau \text{ such that } \tau(g_N) = U_{kl}(g_N^*), 1 \le k, l \le K_0\},$$

where $U_{k,l}(g_N^*)$ merges g_N^* with labels k and l. Furthermore, suppose g_N' gives the unique maximum (up to a permutation τ), and for all H_{g_N}, there exists a positive constant $c_1 > 0$ such that $H_{g_N} \ge 0$, $H_{g_N}^\top \mathbf{1} = \boldsymbol{p}$,

$$\frac{\partial G\{(1-\epsilon)H_{g_N'} + \epsilon H_{g_N}, B^*\}}{\partial \epsilon}\bigg|_{\epsilon=0^+} < -c_1 < 0.$$

For subsampled adjacency matrix $A^{\mathcal{S}}$, consider $\|A^{\mathcal{S}}\|_\infty = \max_{1\le i\le N} \sum_{1\le j\le n} |A_{ij}^{\mathcal{S}}|$. The following Lemma 4 provides a concentration inequality to bound the variation in the adjacency matrix $A^{\mathcal{S}}$, as proposed by [78].

Lemma 4 (Concentration inequality)**.** *Assume $g_N \in \mathcal{C}(A^{\mathcal{S}}, K)$ and define $W_{g_N} = o_{g_N}/M - H_{g_N}B^*H_{g_N}^\top$. For $\epsilon \le 3$,*

$$P\Big\{\max_{g_N\in\mathcal{C}(A^{\mathcal{S}},K)} \|W_{g_N}\|_\infty > \epsilon\Big\} \le 2K^{N+2}\exp\{-c_1(B^*)\epsilon^2\rho_N^{-1}M\},$$

where $c_1(B^*)$ *is a constant depending on* B^* *and* $M = Nn - n(n+1)/2$. *Let* $\omega_n = (\rho_N N \log n/M)^{1/2}$, *then* $\max_{g_N \in \mathcal{C}(A^{\mathcal{S}},K)} \|W_{g_N}\|_\infty > \omega_n \to 0$, *with high probability, for* $n, N \to \infty$. *Furthermore, let* $g'_N \in \mathcal{C}(A^{\mathcal{S}}, K)$ *be a fixed set of labels; then, for* $\epsilon \le \frac{3m}{N}$,

$$P\Big(\max_{g_N : |g_N - g'_N| \le m} \|W_{g_N} - W_{g'_N}\|_\infty > \epsilon\Big) \le 2\binom{N}{m} K^{m+2} \exp\Big\{-c_2(B^*)\frac{N^3\epsilon^2}{\rho_N m}\Big\},$$

where m *is an integer and* $c_2(B^*)$ *is a constant depending on* B^*.

Appendix B: Demonstrations of SM-BIC

In Appendix B, we use the BIC approximation to prove Lemma 1, shown in Appendix B.1. Furthermore, we provide the proofs of Propositions 1 and 2 in Appendices B.2 and B.3, respectively.

B.1. Proof of Lemma 1

The proof of the log-likelihood function approximation can be accomplished by the following two steps. First, we use Taylor approximation for the likelihood function, i.e., $f(A^{\mathcal{S}}|g_N)$. Then, we investigate its Hessian matrix.

Step 1. Assume that the likelihood function $f(A^{\mathcal{S}}|g_N, \theta)$ attains its maximum at $\widehat{\theta}$ so that $\partial f(A^{\mathcal{S}}|g_N,\theta)/\partial\theta|_{\theta=\widehat{\theta}} = 0$. By Taylor expansion, we have,

$$\log f(A^{\mathcal{S}}|g_N,\theta) \approx \log f(A^{\mathcal{S}}|g_N,\widehat{\theta}) + \frac{1}{2}(\theta-\widehat{\theta})^\top D(\theta-\widehat{\theta}),$$

where D is a $\frac{K(K+1)}{2} \times \frac{K(K+1)}{2}$ matrix such that for $1 \le k', l' \le \frac{K(K+1)}{2}$,

$$D_{k'l'} = \frac{\partial^2 f(A^{\mathcal{S}}|g_N,\theta)}{\partial\theta_{k'}\partial\theta_{l'}}\bigg|_{\theta=\hat{\theta}}.$$

Since $f(A^{\mathcal{S}}|g_N,\theta)$ attains its maximum at $\widehat{\theta}$, the Hessian matrix D is negative definite. Let $\widetilde{D} = -D$, and then we approximate $f(A^{\mathcal{S}}|g_N)$,

$$\begin{aligned} f(A^{\mathcal{S}}|g_N) &= \int \exp\{\log f(A^{\mathcal{S}}|g_N,\theta)\}p(\theta)d\theta, \\ &\approx \exp\{\log f(A^{\mathcal{S}}|g_N,\widehat{\theta})\} \times \int \exp\Big\{-\frac{1}{2}(\theta-\widehat{\theta})^\top \widetilde{D}(\theta-\widehat{\theta})\Big\}d\theta. \end{aligned} \tag{B.1}$$

Recognizing the integrand in equation (B.1) as proportional to a multivariate normal density gives

$$f(A^{\mathcal{S}}|g_N) \approx \exp\{\log f(A^{\mathcal{S}}|g_N,\widehat{\theta})\} \times (2\pi)^{\frac{K(K+1)}{4}} |\widetilde{D}|^{-1/2}. \tag{B.2}$$

The use of equation (B.2) is called the *Laplace method for integrals* [64]. The error in equation (B.2) is $O(M^{-1})$ [71, 64], and so

$$\begin{aligned}\log f(A^{\mathcal{S}}|g_N) = &\log f(A^{\mathcal{S}}|g_N,\widehat{\theta}) + \log p(\hat{\theta}) \\ &+ \frac{K(K+1)}{4}\log 2\pi - 1/2\log|\widetilde{D}| + O(M^{-1}).\end{aligned} \tag{B.3}$$

Step 2. To obtain the approximation of the likelihood function, we further study the determinant of $\widetilde{D}$. Recall that the number of independent observations in $A^{\mathcal{S}}$ is M. Let $\{y_r\}_{r=1}^{M}$ denote these independent observations. Then, $\log f(A^{\mathcal{S}}|g_N,\theta) = \sum_{r=1}^{M}\log f(y_r|g_N,\theta)$. Note that

$$\begin{aligned}\widetilde{D}_{k'l'} &= -\frac{\partial^2 \log f(A^{\mathcal{S}}|g_N,\theta)}{\partial\theta_{k'}\theta_{l'}}\bigg|_{\theta=\widehat{\theta}} = -\frac{\partial^2 \log\{\prod_{r=1}^{M} f(y_r|g_N,\theta)\}}{\partial\theta_{k'}\partial\theta_{l'}}\bigg|_{\theta=\widehat{\theta}} \\ &= -\frac{\partial^2\{1/M\sum_{r=1}^{M} M\log f(y_r|g_N,\theta)\}}{\partial\theta_{k'}\theta_{l'}}\bigg|_{\theta=\widehat{\theta}}.\end{aligned}$$

As M grows large, we use the weak law of large numbers on random variables, $x_r = M\log f(y_r|g_N,\theta),\ r=1,\ldots,M$. We obtain

$$1/M\sum_{r=1}^{M} M\log f(y_r|g_N,\theta) \to E\left\{M\log f(y_r|g_N,\theta)\right\},$$

with high probability. Therefore, every element in the observed Fisher information matrix is

$$\widetilde{D}_{k'l'} = -\frac{\partial^2 E\{M\log f(y_r|g_N,\theta)\}}{\partial\theta_{k'}\partial\theta_{l'}}\bigg|_{\theta=\widehat{\theta}} = -M\frac{\partial^2 E\{\log f(y_r|g_N,\theta)\}}{\partial\theta_{k'}\partial\theta_{l'}}\bigg|_{\theta=\widehat{\theta}} = M\widetilde{I}_{k'l'}$$

where $\widetilde{I}_{k'l'}$ is the (k',l')-entry of the Fisher matrix $\widetilde{I}_\theta$ for a single observed y_r $(1\le r\le M)$. Thus,

$$|\widetilde{D}| = (M)^{K(K+1)/2}|\widetilde{I}_\theta|. \tag{B.4}$$

To this end, according to (B.3) and (B.4), we obtain

$$\begin{aligned}\log\{f(A^{\mathcal{S}}|g_N)\} = &\log f(A^{\mathcal{S}}|g_N,\hat{\theta}) + \log p(\hat{\theta}) + \frac{K(K+1)}{4}\log 2\pi \\ &- \frac{K(K+1)}{4}\log M - \frac{1}{2}\log|\widetilde{I}_\theta| + O(M^{-1}).\end{aligned} \tag{B.5}$$

The first term on the right-hand side of (B.5) is of order $O(M)$, the forth term is of order $O(\log M)$, while the other four therms are of order $O(1)$ or less. Thus gives

$$\log\{f(A^{\mathcal{S}}|g_N)\} = \log\{f(A^{\mathcal{S}}|g_N,\widehat{\theta})\} - \frac{K(K+1)}{4}\log M + O(1).$$

This accomplishes the proof.

B.2. Proof of Proposition 1

To demonstrate the effectiveness of our SM-BIC, we first prove the statement regarding the computational complexity of the SM-BIC in Proposition 1. Since the DCSBM is a generalization of the SBM, we discuss the computational complexity of the SM-BIC for DCSBM. According to the SM-BIC, there are two main procedures for determining the number of communities, including node-pair subsampling and the model selection algorithm. Therefore, we analyze the computational complexity of each procedure in detail.

First, the node-pair subsampling procedure includes two steps, where the time complexity of collecting the node set $\mathcal{S}$ is $O(N)$ according to [72], and that of forming an $N \times n$ subsampled adjacency matrix is no more than $O(Nn)$. In this way, the computational complexity of the network subsampling procedure is $O(Nn)$.

Second, perform the model selection algorithm to identify K_0 for the DCSBM. For each candidate K, the SM-BIC evaluates K by the following steps.

(1) Perform spectral clustering to the subsampled adjacency matrix $A^{\mathcal{S}}$ using a truncated SVD, which takes $O(Nn)$ time complexity [28, 55].
(2) Compute the plug-in estimator of B, which requires $O(Nn)$ computational complexity.
(3) Obtain the plug-in estimator of ψ, which has a computational cost of $O(Nn)$.
(4) Calculate the SM-BIC of K with $O(Nn)$.

After repeating steps (1)–(4) $K_{\max}$ times, we obtain the optimal choice of the number of communities.

Since $K_{\max}$ is a constant, the time complexity of the SM-BIC is $O(Nn)$. Therefore, we have proved Proposition 1.

B.3. Proof of Proposition 2

In this section, we accomplish the proof of Proposition 2 by the following two steps. Under the assumptions in Proposition 2, we first prove that the selected node set $\mathcal{S}$ covers K_0 blocks completely with high probability. Then, we demonstrate that the expected average degree of the subnetwork could be $\mathbb{E}(d) = \Omega(\log N)$ with high probability.

Step 1. We first represent the event $\mathcal{S} \in \mathcal{M}_{K_0}$ using some simple events. Specifically, we describe the event $e = \{\mathcal{S} : \ \forall\ k \in [K_0],\ \exists\ i \in \mathcal{S}, g^*_{N,i} = k\}$ using several simple events to simply calculate its probability. Denote $e_k = \{\mathcal{S} : \sum_{i\in\mathcal{S}} \mathbb{I}(g^*_{N,i} = k) > 0\}$, for $k = 1, \ldots, K_0$. Then, we have $e = \bigcap_{k=1}^{K_0} e_k$.

Then, we focus on calculating the probability of event e. Let e^c denote the complement set of e. Then, following De Morgan's laws, $e^c = \bigcup_{k=1}^{K_0} e_k^c$. Therefore, by the property of probability measure,

$$P(e^c) \le \sum_{k=1}^{K_0} P(e_k^c). \tag{B.6}$$

Considering random simple sampling without replacement, the probability of choosing a node from the k-th block is $N_{k,g_N^*}/N$ in each sampling. Then, $P(e_k^c) = (1 - N_{k,g_N^*}/N)^n$, for $k = 1, \ldots, K_0$. As a result, according to (B.6), $P(e^c) \leq \sum_{k=1}^{K_0} P(e_k^c) \leq K_0(1 - N_{\min,g_N^*}/N)^n$. That is, $P(e) > 1 - K_0(1 - N_{\min,g_N^*}/N)^n$, where $N_{\min,g_N^*} = \min_k N_{k,g_N^*}$.

Consider the subsample size n such that $\epsilon \geq K_0(1 - N_{\min,g_N^*}/N)^n$, and then, $n \geq \log(K_0/\epsilon)/\log\{(1 - N_{\min}/N)^{-1}\}$. Under Assumption (A2), we can find a constant c_0 such that $N_{\min,g_N^*}/N > c_0/K_0$. As a result, the subsample size $n \geq \log(K_0/\epsilon)/\log\{K_0/(K_0 - c_0)\}$. Taking $\epsilon = 1/N$ and $K_0 = O(1)$, we have $n = \Omega(\log N)$. Therefore, according to the assumptions in Proposition 2, we have proved $\mathcal{S} \in \mathcal{M}_{K_0}$ with probability $1 - 1/N$.

Step 2. Consider that the network density is ρ_N and under Assumptions (A1)–(A3), we have $\mathbb{E}(d) = \mathbb{E}\{\sum_{i=1}^N \sum_{j=1}^n A_{ij}^{\mathcal{S}}/N\} = \Omega(n\rho_N)$. Furthermore, since $n = \Omega(\log N/\rho_N)$, we have $\mathbb{E}(d) = \Omega(\log N)$. Hence, we proved Proposition 2.

Appendix C: Theoretical proof of SM-BIC

Here, we first establish the consistency of the SM-BIC under the SBM. Specifically, we demonstrate the claim of Theorem 1 in Appendix C.1, and further give the proof of Theorems 2 and 3 in Appendices C.2 and C.3, respectively. Then, we discuss the theoretical property of the SM-BIC under the DCSBM, i.e., Theorem 4, in Appendix C.4.

C.1. Proof of Theorem 1

Without loss of generality, we start with $K = K_0 - 1$. To prove Theorem 1, we focus on analyzing the log-likelihood ratio L_{K_0-1,K_0}, where $L_{K_0-1,K_0} = \max_{g_N \in \mathcal{C}(A^{\mathcal{S}},K_0-1)} \sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}}|g_N, B) - \log f(A^{\mathcal{S}}|g_N^*, B^*)$. Specifically, we accomplish the proof by following three steps. We first analyze the node assignments obtained by $K_0 - 1$ in detail, and then we discuss the likelihood function of L_{K_0-1,K_0}. Finally, we establish the upper bound for L_{K_0-1,K_0}.

Step 1. We discuss the community assignments based on SBM_K. In the under-fitting case, $K = K_0 - 1$, we define a merge mechanism. First, we give the merged label vector set. Define $e_{K_0-1} = \{g_N \in \mathcal{C}(A^{\mathcal{S}}, K_0 - 1) : g_N = U_{k,l}(g_N^*), 1 \leq k \neq l \leq K_0\}$. Therefore, the assignments in e_{K_0-1} merge two blocks in g_N^* into a block. By Lemma 3, without loss of generality, assume that the maximum of $G(H_{g_N}, B^*)$ is achieved at $g_N' = U_{K_0-1,K_0}(g_N^*)$. Then, we establish the corresponding merged connectivity matrix $B' \in \mathbb{B}_{K_0-1}$. Define $U_{k,l}(g_N^*, B^*)$ to represent merging blocks k and l in B^* by taking weighted averages with $\boldsymbol{p}$. Specifically, if $B' = U_{K_0-1,K_0}(g_N^*, B^*)$, then

$$B'_{u(k)u(l)} = \begin{cases} B_{kl}^*, \ 1 \leq k \leq l \leq K_0 - 2; \\ \dfrac{n_{kK_0-1,g_N^*} B_{kK_0-1}^* + n_{kK_0,g_N^*} B_{kK_0}^*}{n_{kK_0-1,g_N^*} + n_{kK_0,g_N^*}}, \ (1 \leq k \leq K_0 - 2, \\ \qquad\qquad\qquad\qquad K_0 - 1 \leq l \leq K_0); \end{cases}$$

where $B'_{u(k)u(l)} = B'_{u(l)u(k)}$ for $1 \le k \le l \le K_0 - 2$. Let $\bar{O}_{K_0-1K_0-1,g^*} = n_{K_0-1K_0-1,g_N^*} B^*_{K_0-1K_0-1}$, $\bar{O}_{K_0-1K_0,g_N^*} = n_{K_0-1K_0,g_N^*} B^*_{K_0-1K_0}$, $\bar{O}_{K_0K_0-1,g_N^*} = n_{K_0K_0-1,g_N^*} B^*_{K_0K_0-1}$, and $\bar{O}_{K_0K_0,g_N^*}$. Then, for $K_0 - 1 \le k, l \le K_0$,

$$B'_{u(k)u(l)} = \frac{\bar{O}_{K_0-1K_0-1,g_N^*} + \bar{O}_{K_0-1K_0,g_N^*} + \bar{O}_{K_0K_0-1,g_N^*} + \bar{O}_{K_0K_0,g_N^*}}{n_{K_0-1K_0-1,g_N^*} + n_{K_0-1K_0,g_N^*} + n_{K_0K_0-1,g_N^*} + n_{K_0K_0,g_N^*}},$$

where $1 \le u(k) \le K_0 - 1$ and $1 \le u(l) \le K_0 - 1$ are the new block labels of communities k and l, respectively.

Step 2. We now study the log-likelihood ratio L_{K_0-1,K_0}. We demonstrate the following critical equation in the first step:

$$\max_{g_N \in \mathcal{C}(A^{\mathcal{S}}, K_0-1)} \sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}} | g_N, B) = \sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}} | g'_N, B). \tag{C.1}$$

The proof of (C.1) can be accomplished in two steps. We first prove this by considering g_N far away from g_N^* and close to g'_N (up to permutation τ). Specifically, define $\mathcal{J}_{\delta_n}^- = \{g_N \in \mathcal{C}(A^{\mathcal{S}}, K_0 - 1) : G(H_{g_N}, B^*) - G(H_{g'_N}, B^*) < -\delta_n\}$, where $\delta_n \to 0$ slowly. Then, we apply some useful lemmas provided earlier to prove this in another case.

Step 2.1. For $g_N \in \mathcal{J}_{\delta_n}^-$, we prove the equality (C.1). By Lemma 4, there exists a constant c_1 such that

$$\begin{aligned} &\left| F\Big(\frac{o_{g_N}}{M}, \frac{n_{g_N}}{M}\Big) - G(H_{g_N}, B^*) \right| \\ &\le c_1 \sum_{1 \le k \le l \le K_0 - 1} \left| \frac{o_{kl,g_N}}{M} - \{H_{g_N} B^* H_{g_N}^\top\}_{kl} \right| = O_P(\omega_n), \end{aligned}$$

where the inequality holds because $\gamma(\cdot)$ is Lipschitz on any interval bounded away from 0 and 1, and recall that $\omega_n = (\rho_N N \log n / M)^{1/2}$. Then, for any $g_N \in \mathcal{J}_{\delta_n}^-$, we have

$$\begin{aligned} &\sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}} | g_N, B) \\ &= \sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}} | g'_N, B) + M\Big\{ F(o_{g_N}/M, n_{g_N}/M) - G(H_{g_N}, B^*) \Big\} \\ &\quad + M\Big\{ G(H_{g_N}, B^*) - G(H_{g'_N}, B^*) \Big\} + M\Big\{ G(H_{g'_N}, B^*) - F\{o_{g'_N}/M, n_{g'_N}/M\} \Big\} \\ &= \sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}} | g'_N, B) + O_P(M\omega_n - M\delta_n + M\omega_n) \end{aligned} \tag{C.2}$$

Hence, we obtain

$$\max_{g_N \in \mathcal{J}_{\delta_n}^-} \sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}} | g_N, B)$$

$$\le \log\Big\{ \sum_{g_N \in \mathcal{J}_{\delta_n}^-} \sup_{B \in \mathbb{B}_{K_0-1}} f(A^{\mathcal{S}}|g_N, B) \Big\}$$

$$= \log\Big[\sum_{g_N \in \mathcal{J}_{\delta_n}^-} \sup_{B \in \mathbb{B}_{K_0-1}} \exp\{\log f(A^{\mathcal{S}}|g_N, B)\} \Big]$$

$$\le \log\Big[\sup_{B \in \mathbb{B}_{K_0-1}} f(A^{\mathcal{S}}|g'_N, B)(K_0-1)^n \exp\{O_P(2M\omega_n - M\delta_n)\} \Big] \tag{C.3}$$

$$\le \log \sup_{B \in \mathbb{B}_{K_0-1}} f(A^{\mathcal{S}}|g'_N, B), \tag{C.4}$$

where (C.3) is derived from (C.2), and if $\delta_n \to 0$ slowly enough such that $\delta_n/\omega_n \to \infty$, we have (C.4).

Step 2.2. For $g_N \notin \mathcal{J}_{\delta_n}^-$, $|G(H_{g_N}, B^*) - G(H_{g'_N}, B^*)| \to 0$. Let $\bar{g}_N := \min_\tau |\tau(g_N) - g'_N|$. Since the maximum is unique up to τ, $\|H_{\bar{g}_N} - H_{g'_N}\|_\infty \to 0$. By Lemma 4,

$$P\left(\max_{g_N \in \tau(g'_N)} \|W_{\bar{g}_N} - W_{g'_N}\|_\infty > \epsilon |\bar{g}_N - g'_N|/N \right)$$

$$\le \sum_{m=1}^{N} P\left(\max_{g_N : g_N = \bar{g}_N, |\bar{g}_N - g'_N| = m} \|W_{\bar{g}_N} - W_{g'_N}\|_\infty > \frac{\epsilon m}{N} \right)$$

$$\le \sum_{m=1}^{N} \left\{ 2(K_0-1)^{K_0-1} N^m (K_0-1)^{m+2} \exp\left(-c_1 \rho_N^{-1} N m\right) \right\} \to 0.$$

It follows for $|\bar{g}_N - g'_N| = m$, $g_N \notin \mathcal{J}_{\delta_n}^-$,

$$\left\| \frac{o_{\bar{g}_N}}{M} - \frac{o_{g'_N}}{M} \right\|_\infty = o_P(1)\frac{|\bar{g}_N - g'_N|}{N} + \left\| H_{\bar{g}_N} B^* H_{\bar{g}_N}^\top - H_{g'_N} B^* H_{g'_N}^\top \right\|_\infty$$

$$\ge \frac{m}{N}(c_1 + o_P(\rho_N)).$$

Observe that $\|\frac{o_{g'_N}}{M} - H_{g'_N} B^* H_{g'_N}^\top\|_\infty = o_P(\rho_N)$. By Lemma 4, $\|n_{g'_N}/M - H_{g'_N} \mathbf{1}\mathbf{1}^\top H_{g'_N}^\top\|_\infty = o_P(\rho_N)$. Note that $F(\cdot,\cdot)$ has a continuous derivative in the neighborhood for $(o_{g'_N}/M, n_{g'_N}/M)$. By Lemma 3,

$$\left. \frac{\partial F\{(1-\epsilon)o_{g'_N}/M + \epsilon Q, (1-\epsilon)n_{g'_N}/M + \epsilon q\}}{\partial \epsilon} \right|_{\epsilon = 0^+} < -c_1 \rho_N < 0,$$

for (Q, q) in the neighborhood of $(o_{g'_N}/M, n_{g'_N}/M)$. Hence, $F\left(o_{\bar{g}_N}/M, n_{\bar{g}_N}/M\right) - F\left(o_{g'_N}/M, n_{g'_N}/M\right) \le -c_1 \rho_N m/N$. Furthermore, we obtain

$$\begin{aligned} &\sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}}|\bar{g}_N, B) - \sup_{B \in \mathbb{B}_{K_0-1}} \log f(A^{\mathcal{S}}|g'_N, B) \\ &= M\left\{ F\left(\frac{o_{\bar{g}_N}}{M}, \frac{n_{\bar{g}_N}}{M}\right) - F\left(\frac{o_{g'_N}}{M}, \frac{n_{g'_N}}{M}\right) \right\} \le -c_1 \frac{m \rho_N M}{N}. \end{aligned} \tag{C.5}$$

Then, we conclude as follows:

$$\begin{aligned}
&\max_{g_N\notin\mathcal{J}_{\delta_n}^-, g_N\notin\tau(g_N')}\sup_{B\in\mathbb{B}_{K_0-1}}\log f(A^{\mathcal{S}}|g_N,B)\\
&\le\log\left\{\sum_{g_N\notin\mathcal{J}_{\delta_n}^-, g_N\notin\tau(g_N')}\sup_{B\in\mathbb{B}_{K_0-1}}\log f(A^{\mathcal{S}}|g_N,B)\right\}\\
&\le\log\left\{\sum_{g_N\in\tau(g_N')}\sup_{B\in\mathbb{B}_{K_0-1}}f(A^{\mathcal{S}}|g_N,B)\sum_{m=1}^{N}\frac{(K_0-1)^mN^m}{\exp(c_1m\rho_NM/N)}\right\} \qquad \text{(C.6)}\\
&\le\log\left[\sup_{B\in\mathbb{B}_{K_0-1}}f(A^{\mathcal{S}}|g_N',B)\sum_{g_N\in\tau(g_N')}\left\{\sum_{m=1}^{N}\frac{(K_0-1)^mN^m}{\exp(c_1m\rho_NM/N)}\right\}\right]\\
&=\sup_{B\in\mathbb{B}_{K_0-1}}\log f(A^{\mathcal{S}}|g_N',B)+\log\left\{(K_0-1)^{K_0-1}\sum_{m=1}^{N}\frac{(K_0-1)^mN^m}{\exp(c_1m\rho_NM/N)}\right\} \qquad \text{(C.7)}\\
&\le\sup_{B\in\mathbb{B}_{K_0-1}}\log f(A^{\mathcal{S}}|g_N',B)+\log\left\{(K_0-1)^{K_0}N^2\exp(-c_1\rho_NM/N)\right\}\\
&=\sup_{B\in\mathbb{B}_{K_0-1}}\log f(A^{\mathcal{S}}|g_N',B)+K_0\log(K_0-1)+2\log N-c_1\rho_NM/N\\
&<\sup_{B\in\mathbb{B}_{K_0-1}}\log f(A^{\mathcal{S}}|g_N',B), \qquad \text{(C.8)}
\end{aligned}$$

where (C.6) is obtained by (C.5), and the equality (C.7) holds because the number of all community assignments in $\tau(g_N')$ is $(K_0-1)^{K_0-1}$. Additionally, the equality (C.8) results from $M/N=\Omega(n)=\Omega(\log N/\rho_N)$. Therefore, by (C.4) and (C.8), we have accomplished the proof of (C.1).

Step 3. We then use the conclusion in (C.1) to give the lower bound of L_{K_0-1,K_0}. We start by analyzing the bias of the maximum likelihood estimator of the connectivity matrix elements. Consider that $\sup_{B\in\mathbb{B}_{K_0-1}}f(A^{\mathcal{S}}|g_N',B)$ is uniquely maximized at

$$\widehat{B}_{kl}=\frac{o_{kl,g_N'}}{n_{kl,g_N'}}=\frac{o_{kl,g_N^*}}{n_{kl,g_N^*}}=B_{kl}^*+O_P(\rho_NM^{-1/2}),\ \text{for } 1\le k\le l\le K_0-2, \qquad \text{(C.9)}$$

$$\widehat{B}'_{kK_0-1}=\frac{o_{kK_0-1,g_N^*}+o_{kK_0,g_N^*}}{n_{kK_0-1,g_N^*}+n_{kK_0,g_N^*}}=B'_{kK_0-1}+O_P(\rho_NM^{-1/2}),\ \text{for } 1\le k\le K_0-2, \qquad \text{(C.10)}$$

$$\widehat{B}'_{K_0-1K_0-1}=\frac{\sum_{k=K_0-1}^{K_0}\sum_{l=K_0-1}^{K_0}o_{kl,g_N^*}}{\sum_{k=K_0-1}^{K_0}\sum_{l=K_0-1}^{K_0}n_{kl,g_N^*}}=B'_{K_0-1K_0-1}+O_P(\rho_NM^{-1/2}), \qquad \text{(C.11)}$$

where the equalities (C.9), (C.10), and (C.11) are derived by Hoeffding's inequality [38] presented in Lemma 2. Hence, we have

$$L_{K_0-1K_0}=\sup_{B\in\mathbb{B}_{K_0-1}}\log f(A^{\mathcal{S}}|g_N',B)-\log f(A^{\mathcal{S}}|g_N^*,B^*)$$

$$= \sum_{1\le k\le l\le K_0-2} \left\{ o_{kl,g_N^*} \log\left(\frac{\widehat{B}_{kl}}{B_{kl}^*}\right) + (n_{kl,g_N^*} - o_{kl,g_N^*}) \log\left(\frac{1-\widehat{B}_{kl}}{1-B_{kl}^*}\right)\right\}$$
$$+ \sum_{k,l\in\mathcal{I}} \left\{ o_{kl,g_N^*} \log\left(\frac{\widehat{B}'_{u(k)u(l)}}{B_{kl}^*}\right) + (n_{kl,g_N^*} - o_{kl,g_N^*}) \log\left(\frac{1-\widehat{B}'_{u(k)u(l)}}{1-B_{kl}^*}\right)\right\}$$

where $\mathcal{I}$ is the set of indices affected by the merge, $\mathcal{I} = \{(k,l) \in [K_0]^2, K_0 - 1 \le l \le K_0, k \le l\}$. For convenience, let

$$X_1 = \sum_{1\le k\le l\le K_0-2} \left\{ o_{kl,g_N^*} \log\left(\frac{\widehat{B}_{kl}}{B_{kl}^*}\right) + (n_{kl,g_N^*} - o_{kl,g_N^*}) \log\left(\frac{1-\widehat{B}_{kl}}{1-B_{kl}^*}\right)\right\}, \tag{C.12}$$

$$X_2 = \sum_{k,l\in\mathcal{I}} \left\{ o_{kl,g_N^*} \log\left(\frac{\widehat{B}'_{u(k)u(l)}}{B_{kl}^*}\right) + (n_{kl,g_N^*} - o_{kl,g_N^*}) \log\left(\frac{1-\widehat{B}'_{u(k)u(l)}}{1-B_{kl}^*}\right)\right\}, \tag{C.13}$$

where X_1 represents the bias within un-merged communities (i.e., $1 \le k, l \le K_0-2$) and X_2 measure the bias within the merged communities (i.e., $(k,l) \in \mathcal{I}$). That is $L_{K_0-1K_0} = X_1 + X_2$. Next, we discuss X_1 and X_2, accordingly.

First, by Taylor's expansion, we obtain

$$\begin{aligned} X_1 &= \sum_{1\le k\le l\le K_0-2} \left\{ o_{kl,g_N^*} \log\left(\frac{\widehat{B}_{kl}}{B_{kl}^*}\right) + (n_{kl,g_N^*} - o_{kl,g_N^*}) \log\left(\frac{1-\widehat{B}_{kl}}{1-B_{kl}^*}\right)\right\} \\ &= \sum_{1\le k\le l\le K_0-2} \left[n_{kl,g_N^*}(B_{kl}^* + \Delta_{kl}) \left\{ \frac{\Delta_{kl}}{B_{kl}^*} - \frac{\Delta_{kl}^2}{2(B_{kl}^*)^2} \right\} \right. \\ &\quad \left. + n_{kl,g_N^*}(1 - B_{kl}^* - \Delta_{kl}) \left\{ \frac{-\Delta_{kl}}{1-B_{kl}^*} - \frac{\Delta_{kl}^2}{2(1-B_{kl}^*)^2} \right\} + O(n_{kl,g_N^*}\Delta_{kl}^3) \right] \end{aligned} \tag{C.14}$$

$$\begin{aligned} &= \sum_{1\le k\le l\le K_0-2} \left[n_{kl,g_N^*} \left(\Delta_{kl} + \frac{\Delta_{kl}^2}{2B_{kl}^*} \right) \right. \\ &\quad \left. + n_{kl,g_N^*} \left\{ -\Delta_{kl} + \frac{\Delta_{kl}^2}{2(1-B_{kl}^*)} \right\} + O(n_{kl,g_N^*}\Delta_{kl}^3) \right] \\ &= \frac{1}{2} \sum_{1\le k\le l\le K_0-2} \frac{n_{kl,g_N^*}(\widehat{B}_{kl} - B_{kl}^*)^2}{B_{kl}^*(1-B_{kl}^*)} + O_P(\rho_N^3 M^{-1/2}), \end{aligned} \tag{C.15}$$

where $\Delta_{kl} = \widehat{B}_{kl} - B_{kl}^*$ in equality (C.14), and (C.15) results from (C.9). Hence, the upper bound of (C.12) is $O_P(\rho_N)$. Then, we focus on (C.13). By Taylor

expansion, we have

$$\begin{aligned} X_2 &= \sum_{k,l\in\mathcal{I}} \left[o_{kl,g_N^*} \log\left\{ \frac{\widehat{B}'_{u(k)u(l)}}{B^*_{kl}} \right\} + (n_{kl,g_N^*} - o_{kl,g_N^*}) \log\left\{ \frac{1-\widehat{B}'_{u(k)u(l)}}{1-B^*_{kl}} \right\} \right] \\ &= \sum_{k,l\in\mathcal{I}} n_{kl,g_N^*} \left[B'_{u(k)u(l)} \log\left\{ \frac{B'_{u(k)u(l)}(1-B^*_{kl})}{(1-B'_{u(k)u(l)})B^*_{kl}} \right\} \right. \\ &\quad \left. + \log\left\{ \frac{1-B'_{u(k)u(l)}}{1-B^*_{kl}} \right\} + O(\Delta'_{u(k)u(l)}) \right] \qquad \text{(C.16)} \\ &= -\Omega_P(\rho_N M), \end{aligned}$$

where $\Delta'_{u(k)u(l)} = \widehat{B}'_{u(k)u(l)} - B'_{u(k)u(l)}$. Hence, we have $L_{K_0-1,K_0} = X_1 + X_2 = -\Omega_P(\rho_N M)$. Therefore, we have accomplished the proof of Theorem 1.

C.2. *Proof of Theorem 2*

Based on the proof of Theorem 1, we prove the convergence of the penalized log-likelihood function $\ell(K_0)$ via the following two steps.

STEP 1. For $K = K_0$, according to (C.1), we have

$$\max_{g_N\in\mathcal{C}(A^{\mathcal{S}},K_0)} \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N,B) = \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*,B). \qquad \text{(C.17)}$$

Hence, $L_{K_0,K_0} = \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*,B) - \log f(A^{\mathcal{S}}|g_N^*,B^*)$.

STEP 2. Consider that $\sup_{B\in\mathbb{B}_{K_0}} f(A^{\mathcal{S}}|g_N^*,B)$ is uniquely maximized at

$$\widehat{B}_{kl} = \frac{o_{kl,g_N^*}}{n_{kl,g_N^*}} = B^*_{kl} + O_P(\rho_N M^{-1/2}), \text{ for } 1\le k\le l\le K_0. \qquad \text{(C.18)}$$

Then, similar to (C.14), by Taylor expansion, we have,

$$\begin{aligned} &\sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*,B) - \log f(A^{\mathcal{S}}|g_N^*,B^*) \\ &= \sum_{1\le k\le l\le K_0} \left\{ o_{kl,g_N^*} \log\left\{ \frac{\widehat{B}_{kl}(1-B^*_{kl})}{B^*_{kl}(1-\widehat{B}_{kl})} \right\} + n_{kl,g_N^*} \log\left(\frac{1-\widehat{B}_{kl}}{1-B^*_{kl}} \right) + O(n_{kl,g_N^*}\Delta^3_{kl}) \right\} \\ &= \frac{1}{2} \sum_{1\le k\le l\le K_0} \frac{n_{kl,g_N^*}(\widehat{B}_{kl}-B^*_{kl})^2}{B^*_{kl}(1-B^*_{kl})} + O_P(\rho_N^3 M^{-1/2}). \qquad \text{(C.19)} \end{aligned}$$

The last equality results from (C.18), that is $\Delta_{kl} = O_P(\rho_N M^{-1/2})$. Hence, $L_{K_0,K_0} = O_P(\rho_N)$, and this accomplishes the proof of Theorem 2.

C.3. Proof of Theorem 3

Based on the proof of Theorem 2, we define a log-likelihood ratio as

$$\widetilde{L}_{K,K_0} = \max_{g_N \in \mathcal{C}(A^{\mathcal{S}},K)} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N,B) - \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*,B).$$

To provide the upper bound of L_{K,K_0}, we start by discussing $\widetilde{L}_{K,K_0}$. Specifically, we establish the upper bound of L_{K,K_0} by the following three steps. First, we introduce a set of community assignments that is formed by splitting the underlying node assignments into K blocks. Second, we study the corresponding likelihood functions of $\widetilde{L}_{K,K_0}$. Third, based on the conclusion of $\widetilde{L}_{K,K_0}$, we use the preceding lemmas to accomplish this proof.

Step 1. We first define the community assignment set by splitting the underlying g_N^* into K blocks. Intuitively, embedding a K_0-block model in a larger model can be achieved by appropriately splitting the labels g_N^*. Specifically, we define a subset

$$e_K = \{g_N \in \mathcal{C}(A^{\mathcal{S}},K) : \text{each row of } H_{g_N} \text{ has at most one nonzero entry}\}.$$

Then, any $g_N \in e_K$ satisfies the following: every block in g_N is a subset of an existing block in g_N^*. Accordingly, we define a surjective function as $h : [K] \to [K_0]$ describing the assignments in H_{g_N}. In other words, for any $k \in [K]$, $h(k) \in [K_0]$, and $\forall\ a \in [K_0]$, $h^{-1}(a) \in [K]$.

Step 2. We then discuss the log-likelihood ratio L_{K,K_0}. Note that, in this case, $G(H_{g_N},B^*)$ is maximized at any $g_N' \in e_K$ with value $\sum_{1\le k\le l\le K_0} \boldsymbol{p}_k\boldsymbol{p}_l\gamma(B^*_{kl})$. Denote the optimal $G^* = \sum_{1\le k\le l\le K_0} \boldsymbol{p}_k\boldsymbol{p}_l\gamma(B^*_{kl})$. Let $\mathcal{J}^+_{\delta_n} = \{g_N \in \mathcal{C}(A^{\mathcal{S}},K) : G(H_{g_N},B^*) - G^* < -\delta_n\}$, for $\delta_n \to 0$ slowly enough. Then, to analyze the log-likelihood ratio $\widetilde{L}_{K,K_0}$, we consider the likelihood $\sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N,B)$ under two cases, namely, the community assignment $g_N \in \mathcal{J}^+_{\delta_n}$ and $g_N \notin \mathcal{J}^+_{\delta_n}$.

Step 2.1 We analyze $\sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N,B)$ by considering $g_N \in \mathcal{J}^+_{\delta_n}$. By Lemma 4, we have

$$\left|F\left(\frac{o_{g_N}}{M},\frac{n_{g_N}}{M}\right) - G(H_{g_N},B^*)\right| \le c_1 \sum_{1\le k\le l\le K} \left|\frac{o_{kl,g_N}}{M} - (H_{g_N}B^*H_{g_N}^\top)_{kl}\right| = O_P(\omega_n).$$

Therefore, for any $g_N' \in e_K$, we obtain

$$\begin{aligned}
&\max_{g_N\in\mathcal{J}^+_{\delta_n}} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N,B)\\
&\le \log\Big\{\sum_{g_N\in\mathcal{J}^+_{\delta_n}} \sup_{B\in\mathbb{B}_K} f(A^{\mathcal{S}}|g_N,B)\Big\}\\
&= \log\Big[\sum_{g_N\in\mathcal{J}^+_{\delta_n}} \sup_{B\in\mathbb{B}_K} \exp\Big\{\log f(A^{\mathcal{S}}|g_N,B)\Big\}\Big]
\end{aligned}$$

$$\le \log\Big[\sup_{B\in\mathbb{B}_K} f(A|g'_N,B)(K-1)^n \exp\{O_P(2M\omega_n - M\delta_n)\}\Big]$$
$$\le \log\Big\{\sup_{B\in\mathbb{B}_K} f(A|g'_N,B)\Big\}$$
$$= \sup_{B\in\mathbb{B}_K} \sum_{1\le a\le b\le K_0} \sum_{(k,l)\in h^{-1}(a)\times h^{-1}(b)} \{o_{kl,g'_N}\log\Big(\frac{B_{kl}}{1-B_{kl}}\Big) + n_{kl,g'_N}\log(1-B_{kl})\}. \tag{C.20}$$

Choosing $\delta_n \to 0$ slowly enough such that $\delta_n/\omega_n \to \infty$.

We further analyze equality (C.20). Let

$$l_{ab} = \sum_{(k,l)\in h^{-1}(a)\times h^{-1}(b)} \Big\{o_{kl,g'_N}\log B_{kl} + (n_{kl,g'_N} - o_{kl,g'_N})\log(1-B_{kl})\Big\} + \lambda'\Big(\sum_{(k,l)\in h^{-1}(a)\times h^{-1}(b)} n_{kl,g'_N} - n_{ab,g^*_N}\Big).$$

Then, $\partial l_{ab}/\partial n_{kl,g'_N} = \log(1-B_{kl}) + \lambda' = 0$. This implies that for $(k,l) \in h^{-1}(a)\times h^{-1}(b)$, B_{kl}'s are all equal. Let $B_{kl} = B_{ab}$. Hence,

$$\sum_{(k,l)\in h^{-1}(a)\times h^{-1}(b)} \{o_{kl,g'_N}\log B_{kl} + (n_{kl,g'_N} - o_{kl,g'_N})\log(1-B_{kl})\}$$
$$= o_{ab,g^*_N}\log B_{ab} + (n_{ab,g^*_N} - o_{ab,g^*_N})\log(1-B_{ab}),$$

where

$$o_{ab,g^*_N} = \sum_{(k,l)\in h^{-1}(a)\times h^{-1}(b)} o_{kl,g'_N} \quad \text{and} \quad n_{ab,g^*_N} = \sum_{(k,l)\in h^{-1}(a)\times h^{-1}(b)} n_{kl,g'_N}.$$

Therefore, based on (C.20), we have

$$\max_{g_N\in\mathcal{J}^+_{\delta_n}} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N,B)$$
$$\le \sup_{B\in\mathbb{B}_K} \sum_{1\le a\le b\le K_0} \sum_{(k,l)\in h^{-1}(a)\times h^{-1}(b)} \Big\{o_{kl,g'_N}\log\Big(\frac{B_{kl}}{1-B_{kl}}\Big) + n_{kl,g'_N}\log(1-B_{kl})\Big\}$$
$$= \sup_{B\in\mathbb{B}_{K_0}} \sum_{1\le a\le b\le K_0} \{o_{ab,g^*_N}\log B_{ab} + (n_{ab,g^*_N} - o_{ab,g^*_N})\log(1-B_{ab})\}$$
$$= \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g^*_N,B). \tag{C.21}$$

Step 2.2 We investigate the likelihood function $\sup_{B\in\mathbb{B}_K}\log f(A^{\mathcal{S}}|g_N,B)$ for $g_N \notin \mathcal{J}^+_{\delta_n}$. Treating H_{g_N} as a vector, $\{H_{g_N} : g_N \in e_K\}$ is a subset of the union of some of the $K-K_0$ faces of polyhedron $P_{H_{g_N}}$. For every $g_N \notin e_K$, $g_N \notin \mathcal{J}^+_{\delta_n}$, let $g_\perp$ be such that $H_{g_\perp} := \min_{H_{g'_N}: g'_N\in e_K} \|H_{g_N} - H_{g'_N}\|_2$. Then, $H_{g_N} - H_{g_\perp}$

is perpendicular to the corresponding $K - K_0$ face. This orthogonal implies that the directional derivative of $G(\cdot, B^*)$ along the direction of $H_{g_N} - H_{g_\perp}$ is bounded away from 0. That is,

$$\frac{\partial G\{(1-\epsilon)H_{g_\perp} + \epsilon H_{g_N}, B^*\}}{\partial \epsilon}\Big|_{\epsilon=0^+} < -c_1\rho_N,$$

for some universal positive constant c_1. Then, similar to the proof of Theorem 1, we obtain $\sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) - \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_\perp, B) \le -c_1 m\rho_N M/N$, for $|g_N - g_\perp| = m$. Hence, we have

$$\begin{aligned}
&\max_{g_N\notin\mathcal{J}^+_{\delta_n}, g_N\notin e_K} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) \\
&\le \max_{g_N\in e_K} \sup_{B\in\mathbb{B}_K} \log\Big\{f(A^{\mathcal{S}}|g_N, B) \times \sum_{m=1}^{N} (K-1)^m N^m \exp\left(-c_1 m\rho_N M/N\right)\Big\}
\end{aligned}$$

That is,

$$\begin{aligned}
&\max_{g_N\notin\mathcal{J}^+_{\delta_n}, g_N\notin e_K} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) \\
&\le \max_{g_N\in e_K} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) + 2\log N + \log K - \frac{c_1\rho_N M}{N}
\end{aligned}$$

Let $\mu_N = 2 + \log K/\log N - c_1\rho_N M/(N\log N)$ and further by (C.20), we have

$$\max_{g_N\notin\mathcal{J}^+_{\delta_n}, g_N\notin e_K} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) \le \mu_N \log N + \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*, B) \tag{C.22}$$

where the inequality (C.22) is obtained by (C.21). To this end, according to (C.21) and (C.22), we have $\widetilde{L}_{K,K_0} \le \mu_N \log N$.

Step 3. Based on the assertion, $\widetilde{L}_{K,K_0} \le \mu_N \log N$, we now bound the divergence of L_{K,K_0}. According to (C.19) in the proof of Theorem 2, we have

$$\begin{aligned}
L_{K,K_0} &= \max_{g_N\in\mathcal{C}(A^{\mathcal{S}},K)} \sup_{B\in\mathbb{B}_K} \log f(A^{\mathcal{S}}|g_N, B) - \log f(A^{\mathcal{S}}|g_N^*, B^*) \\
&\le \mu_N \log N + \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*, B) - \log f(A^{\mathcal{S}}|g_N^*, B^*) \\
&\le \mu_N \log N + \frac{1}{2} \sum_{1\le k\le l\le K_0} \frac{n_{kl,g_N^*}(\widehat{B}_{kl} - B^*_{kl})^2}{B^*_{kl}(1-B^*_{kl})} + O_P(\rho_N^3 M^{-1/2}) \\
&= \mu_N \log N + O_P(\rho_N),
\end{aligned} \tag{C.23}$$

where the last inequality is according to (C.19). Hence, $L_{K,K_0} = O_P(\mu_N \log N)$ where $\mu_N = O_P(1)$ for $n, N \to \infty$. Therefore, we have accomplished the proof of Theorem 3.

C.4. Proof of Theorem 4

Now, we prove the convergence of the log-likelihood ratio for the DCSBM by the following two steps.

STEP 1. For $K = K_0$, according to (C.17), we obtain

$$\begin{aligned} L_{K_0,K_0} &= \max_{g_N \in \mathcal{C}(A^{\mathcal{S}},K_0)} \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N,B,\psi^*) - \log f(A^{\mathcal{S}}|g_N^*,B^*,\psi^*) \\ &= \sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*,B,\psi^*) - \log f(A^{\mathcal{S}}|g_N^*,B^*,\psi^*). \end{aligned} \tag{C.24}$$

Consider that $\sup_{B\in\mathbb{B}_{K_0}} f(A^{\mathcal{S}}|g_N^*,B,\psi^*)$ is uniquely maximized at $\widehat{B}_{kl} = o_{kl,g_N^*}/n_{kl,g_N^*}(\psi^*)$, for $1 \le k \le l \le K_0$. By Lemma 2, for any $s > 0$, we have

$$\begin{aligned} P(|\widehat{B}_{kl} - B_{kl}^*| > s) &= P(|\rho_N(\rho_N^{-1}\widehat{B}_{kl}) - \rho_N\widetilde{B}_{kl}^*| > s) \\ &\le P(|\rho_N^{-1}\widehat{B}_{kl} - \widetilde{B}_{kl}^*| > \rho_N^{-1}s) \\ &\le \exp\{-2\rho_N^{-2}s^2 n_{kl,g_N^*}(\psi^*)\}. \end{aligned}$$

Hence, $\Delta_{kl} = \widehat{B}_{kl} - B_{kl}^* = O_P\{\rho_N n_{kl,g_N^*}^{-1/2}(\psi^*)\}$, for $1 \le k \le l \le K_0$.

STEP 2. Similar to (C.19), by Taylor expansion, we have

$$\begin{aligned} &\sup_{B\in\mathbb{B}_{K_0}} \log f(A^{\mathcal{S}}|g_N^*,B,\psi^*) - \log f(A^{\mathcal{S}}|g_N^*,B^*,\psi^*) \\ &= \frac{1}{2}\sum_{1\le k\le l\le K_0} \Big[\frac{n_{kl,g_N^*}(\psi^*)\Delta_{kl}^2}{B_{kl}^*} + O\{n_{kl,g_N^*}(\psi^*)\Delta_{kl}^3\}\Big]. \end{aligned} \tag{C.25}$$

Since $\Delta_{kl} = O_P\{\rho_N n_{kl,g_N^*}^{-1/2}(\psi^*)\}$, by (C.24) and (C.25), we obtain $L_{K_0,K_0} = O_P(\rho_N)$. Therefore, we have accomplished this proof.

Funding

This research is supported by the MOE Project of Key Research Institute of Humanities and Social Sciences (grant 22JJD110001) and the Public Computing Cloud, Renmin University of China. Danyang Huang's research is partially supported by the National Natural Science Foundation of China (grants 72471230 and 12071477) and the fund for building world-class universities (disciplines) at Renmin University of China. Xiangyu Chang's research is partially supported by National Natural Science Foundation for Outstanding Young Scholars of China (grant 72122018). Bo Zhang's research is partially supported by the National Natural Science Foundation of China (grant 72271232), and the fund for building world-class universities (disciplines) at Renmin University of China.

References

[1] Adamic, L. A. and Glance, N. (2005), "The political blogosphere and the 2004 US election: divided they blog", in *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36–43.

[2] Akcora, C. G., Gel, Y. R., Kantarcioglu, M., Lyubchich, V., and Thuraisingham, B. (2019), "Graphboot: Quantifying uncertainty in node feature learning on large networks," *IEEE Transactions on Knowledge and Data Engineering*, 33, 116–127.

[3] Amini, A. A., Chen, A., Bickel, P. J., Levina, E., et al. (2013), "Pseudo-likelihood methods for community detection in large sparse networks," *The Annals of Statistics*, 41, 2097–2122. MR3127859

[4] Amini, A. A. and Levina, E. (2018), "On semidefinite relaxations for the block model," *The Annals of Statistics*, 46, 149–179. MR3766949

[5] Assadi, S., Kapralov, M., and Khanna, S. (2018), "A simple sublinear-time algorithm for counting arbitrary subgraphs via edge sampling," *arXiv preprint arXiv:1811.07780*. MR3899800

[6] Bamberger, B., Homburg, C., and Wielgos, D. M. (2021), "Wage inequality: Its impact on customer satisfaction and firm performance," *Journal of Marketing*, 85, 24–43.

[7] Bhattacharya, B. B., Das, S., and Mukherjee, S. (2022), "Motif estimation via subgraph sampling: The fourth-moment phenomenon," *The Annals of Statistics*, 50, 987–1011. MR4404926

[8] Bhattacharyya, S. and Bickel, P. J. (2015), "Subsampling bootstrap of count features of networks," *The Annals of Statistics*, 43, 2384–2411. MR3405598

[9] Bickel, P. J. and Chen, A. (2009), "A nonparametric view of network models and Newman–Girvan and other modularities," *Proceedings of the National Academy of Sciences*, 106, 21068–21073.

[10] Bickel, P. J. and Sarkar, P. (2016), "Hypothesis testing for automated community detection in networks," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 253–273. MR3453655

[11] Bodapati, A. V. (2008), "Recommendation systems with purchase data," *Journal of Marketing Research*, 45, 77–93.

[12] Bordenave, C., Lelarge, M., and Massoulié, L. (2015), "Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, IEEE, pp. 1347–1357. MR3473374

[13] Cai, T. T. and Li, X. (2015), "Robust and computationally feasible community detection in the presence of arbitrary outlier nodes," *The Annals of Statistics*, 43, 1027–1059. MR3346696

[14] Cerqueira, A. and Leonardi, F. (2020), "Estimation of the number of communities in the stochastic block model," *IEEE Transactions on Information Theory*, 66, 6403–6412. MR4173547

[15] Chakrabarty, S., Sengupta, S., and Chen, Y. (2025), "Subsampling-based Community Detection for Large Networks," *Statistica Sinica*.

[16] Chaudhuri, K., Chung, F., and Tsiatas, A. (2012), "Spectral clustering of graphs with general degrees in the extended planted partition model," in *Conference on Learning Theory*, pp. 35–1.

[17] Chen, J. and Chen, Z. (2008), "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, 95, 759–771. MR2443189

[18] Chen, K. and Lei, J. (2018), "Network cross-validation for determining the number of communities in network data," *Journal of the American Statistical Association*, 113, 241–251. MR3803461

[19] Chen, S. and Onnela, J.-P. (2019), "A bootstrap method for goodness of fit and model selection with a single observed network," *Scientific Reports*, 9, 1–12.

[20] Chen, X. and Cai, D. (2011), "Large scale spectral clustering with landmark-based representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25.

[21] Dall'Amico, L., Couillet, R., and Tremblay, N. (2021), "A Unified Framework for Spectral Clustering in Sparse Graphs." *Journal of Machine Learning Research*, 22, 217–1. MR4329796

[22] Daudin, J.-J., Picard, F., and Robin, S. (2008), "A mixture model for random graphs," *Statistics and Computing*, 18, 173–183. MR2390817

[23] Deng, J., Ding, Y., Zhu, Y., Huang, D., Jing, B., and Zhang, B. (2021), "Subsampling Spectral Clustering for Large-Scale Social Networks," *arXiv preprint arXiv:2110.13613*.

[24] Deng, J., Huang, D., Ding, Y., Zhu, Y., Jing, B., and Zhang, B. (2024), "Subsampling spectral clustering for stochastic block models in large-scale networks," *Computational Statistics & Data Analysis*, 189, 107835. MR4640220

[25] Ding, Y., Pan, R., Zhang, Y., and Zhang, B. (2023), "A matrix completion bootstrap method for estimating scale-free network degree distribution," *Knowledge-Based Systems*, 277, 110803.

[26] Eden, T., Levi, A., Ron, D., and Seshadhri, C. (2017), "Approximately counting triangles in sublinear time," *SIAM Journal on Computing*, 46, 1603–1646. MR3709896

[27] Feige, U. (2004), "On sums of independent random variables with unbounded variance, and estimating the average degree in a graph," in *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 594–603. MR2121648

[28] Feng, X., Yu, W., and Li, Y. (2018), "Faster matrix completion using randomized SVD," in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, pp. 608–615.

[29] Fortunato, S. (2010), "Community detection in graphs," *Physics Reports*, 486, 75–174. MR2580414

[30] Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018), "Community detection in degree-corrected block models," *The Annals of Statistics*, 46, 2153–2185. MR3845014

[31] Girvan, M. and Newman, M. E. (2002), "Community structure in social and

biological networks," *Proceedings of the National Academy of Sciences*, 99, 7821–7826. MR1908073

[32] Goldreich, O. and Ron, D. (2008), "Approximating average parameters of graphs," *Random Structures & Algorithms*, 32, 473–493. MR2422391

[33] Gonen, M., Ron, D., and Shavitt, Y. (2011), "Counting stars and other small subgraphs in sublinear-time," *SIAM Journal on Discrete Mathematics*, 25, 1365–1411. MR2837605

[34] Good, B. H., De Montjoye, Y.-A., and Clauset, A. (2010), "Performance of modularity maximization in practical contexts," *Physical Review E*, 81, 046106. MR2736215

[35] Green, A. and Shalizi, C. R. (2022), "Bootstrapping exchangeable random graphs," *Electronic Journal of Statistics*, 16, 1058–1095. MR4377133

[36] Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011), "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, 53, 217–288. MR2806637

[37] Hastings, M. B. (2006), "Community detection as an inference problem," *Physical Review E*, 74, 035102.

[38] Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30. MR0144363

[39] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic blockmodels: First steps," *Social Networks*, 5, 109–137. MR0718088

[40] Hsieh, C.-S., Hsu, Y.-C., Ko, S. I., Kovářík, J., and Logan, T. D. (2024), "Non-representative sampled networks: Estimation of network structural properties by weighting," *Journal of Econometrics*, 240, 105689. MR4703367

[41] Hu, J., Qin, H., Yan, T., and Zhao, Y. (2020), "Corrected Bayesian information criterion for stochastic block models," *Journal of the American Statistical Association*, 115, 1771–1783. MR4189756

[42] Hwang, N., Xu, J., Chatterjee, S., and Bhattacharyya, S. (2023), "On the estimation of the number of communities for sparse networks," *Journal of the American Statistical Association*, 1–22. MR4797910

[43] Illenberger, J. and Flötteröd, G. (2012), "Estimating network properties from snowball sampled data," *Social Networks*, 34, 701–711.

[44] Jin, J., Ke, Z. T., Luo, S., and Wang, M. (2023), "Optimal estimation of the number of network communities," *Journal of the American Statistical Association*, 118, 2101–2116. MR4646629

[45] Karrer, B. and Newman, M. E. (2011), "Stochastic blockmodels and community structure in networks," *Physical review E*, 83, 016107. MR2788206

[46] Knuth, D. E. (1976), "Big omicron and big omega and big theta," *ACM Sigact News*, 8, 18–24.

[47] Le, C. M. and Levina, E. (2015), "Estimating the number of communities in networks by spectral methods," *arXiv preprint arXiv:1507.00827*. MR4422967

[48] Lei, J. (2016), "A goodness-of-fit test for stochastic block models," *The Annals of Statistics*, 44, 401–424. MR3449773

[49] Lei, J., Rinaldo, A., et al. (2015), "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, 43, 215–237. MR3285605
[50] Li, M., Lian, X.-C., Kwok, J. T., and Lu, B.-L. (2011), "Time and space efficient spectral clustering via column sampling," in *CVPR 2011*, IEEE, pp. 2297–2304.
[51] Li, T., Levina, E., and Zhu, J. (2020), "Network cross-validation by edge sampling," *Biometrika*, 107, 257–276. MR4108931
[52] Li, W. (2013), "Revealing network communities with a nonlinear programming method," *Information Sciences*, 229, 18–28. MR3018716
[53] Lunde, R. and Sarkar, P. (2023), "Subsampling sparse graphons under minimal assumptions," *Biometrika*, 110, 15–32. MR4565441
[54] Ma, S., Su, L., and Zhang, Y. (2021), "Determining the number of communities in degree-corrected stochastic block models," *Journal of Machine Learning Research*, 22, 1–63. MR4253762
[55] Martin, L., Loukas, A., and Vandergheynst, P. (2018), "Fast approximate spectral clustering for dynamic networks," in *International Conference on Machine Learning*, PMLR, pp. 3423–3432.
[56] Mukherjee, S. S., Sarkar, P., and Bickel, P. J. (2021), "Two provably consistent divide-and-conquer clustering algorithms for large networks," *Proceedings of the National Academy of Sciences*, 118, e2100482118. MR4390029
[57] Newman, M. E. (2006), "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, 103, 8577–8582. MR2676073
[58] Newman, M. E. and Girvan, M. (2004), "Finding and evaluating community structure in networks," *Physical Review E*, 69, 026113. MR2282139
[59] Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002), "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, pp. 849–856.
[60] Nowicki, K. and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087. MR1947255
[61] Pattison, P. E., Robins, G. L., Snijders, T. A., and Wang, P. (2013), "Conditional estimation of exponential random graph models from snowball sampling designs," *Journal of Mathematical Psychology*, 57, 284–296. MR3137882
[62] Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, Springer Science & Business Media. MR1707286
[63] Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019), "Speeding up MCMC by efficient data subsampling," *Journal of the American Statistical Association*, 114, 831–843. MR3963184
[64] Raftery, A. E. (1995), "Bayesian model selection in social research," *Sociological methodology*, 111–163.
[65] Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, 39, 1878–1915. MR2893856
[66] Saldana, D. F., Yu, Y., and Feng, Y. (2017), "How many communities are

there?" *Journal of Computational and Graphical Statistics*, 26, 171–181. MR3610418

[67] Shaddy, F. and Shah, A. K. (2022), "When to use markets, lines, and lotteries: How beliefs about preferences shape beliefs about allocation," *Journal of Marketing*, 86, 140–156.

[68] Snijders, T. A., Borgatti, S. P., et al. (1999), "Non-parametric standard errors and tests for network statistics," *Connections*, 22, 161–170.

[69] Snijders, T. A. and Nowicki, K. (1997), "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *Journal of Classification*, 14, 75–100. MR1449742

[70] Thompson, M. E., Ramirez Ramirez, L. L., Lyubchich, V., and Gel, Y. R. (2016), "Using the bootstrap for statistical inference on random graphs," *Canadian Journal of Statistics*, 44, 3–24. MR3474218

[71] Tierney, L. and Kadane, J. B. (1986), "Accurate approximations for posterior moments and marginal densities," *Journal of the american statistical association*, 81, 82–86. MR0830567

[72] Vitter, J. S. (1985), "Random sampling with a reservoir," *ACM Transactions on Mathematical Software (TOMS)*, 11, 37–57. MR0793056

[73] Von Luxburg, U. (2007), "A tutorial on spectral clustering," *Statistics and Computing*, 17, 395–416. MR2409803

[74] Wang, H. and Ma, Y. (2021), "Optimal subsampling for quantile regression in big data," *Biometrika*, 108, 99–112. MR4226192

[75] Wang, H., Yang, M., and Stufken, J. (2019), "Information-based optimal subdata selection for big data linear regression," *Journal of the American Statistical Association*, 114, 393–405. MR3941263

[76] Wang, H., Zhu, R., and Ma, P. (2018), "Optimal subsampling for large sample logistic regression," *Journal of the American Statistical Association*, 113, 829–844. MR3832230

[77] Wang, J., Zhang, J., Liu, B., Zhu, J., and Guo, J. (2021), "Fast network community detection with profile-pseudo likelihood methods," *Journal of the American Statistical Association*, 0, 1–14. MR4595500

[78] Wang, Y. R. and Bickel, P. J. (2017), "Likelihood-based model selection for stochastic block models," *The Annals of Statistics*, 45, 500–528. MR3650391

[79] Yan, B., Sarkar, P., and Cheng, X. (2018), "Provable estimation of the number of blocks in block models," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1185–1194.

[80] Yan, D., Huang, L., and Jordan, M. I. (2009), "Fast approximate spectral clustering," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 907–916.

[81] Yedidia, J. S., Freeman, W. T., Weiss, Y., et al. (2003), "Understanding belief propagation and its generalizations," *Exploring Artificial Intelligence in the New Millennium*, 8, 236–239.

[82] Yu, J., Wang, H., Ai, M., and Zhang, H. (2022), "Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data," *Journal of the American Statistical Association*, 117, 265–276. MR4399084

[83] Zhang, H., Guo, X., and Chang, X. (2022), "Randomized spectral cluster-

ing in large-scale stochastic block models," *Journal of Computational and Graphical Statistics*, 0, 1–52. MR4495720

[84] Zhang, Y. and Xia, D. (2022), "Edgeworth expansions for network moments," *The Annals of Statistics*, 50, 726–753. MR4404918

[85] Zhao, Y., Levina, E., and Zhu, J. (2011), "Community extraction for social networks," *Proceedings of the National Academy of Sciences*, 108, 7321–7326.

[86] Zhao, Y., Levina, E., and Zhu, J. (2012), "Consistency of community detection in networks under degree-corrected stochastic block models," *The Annals of Statistics*, 40, 2266–2292. MR3059083