Submitted 01/20; Published X

Determining the Number of Communities in Degree-corrected Stochastic Block Models

Shujie Ma
Department of Statistics
University of California
Riverside, CA 92521, USA

Liangjun Su

Editor:

Abstract

We propose to estimate the number of communities in degree-corrected stochastic block models based on a pseudo likelihood ratio statistic. To this end, we introduce a method that combines spectral clustering with binary segmentation. This approach guarantees an upper bound for the pseudo likelihood ratio statistic when the model is over-fitted. We also derive its limiting distribution when the model is under-fitted. Based on these properties, we establish the consistency of our estimator for the true number of communities. Developing these theoretical properties require a mild condition on the average degrees – growing at a rate no slower than $\log(n)$, where n is the number of nodes. Our proposed method is further illustrated by simulation studies and analysis of real-world networks. The numerical results show that our approach has satisfactory performance when the network is semi-dense.

Keywords: Clustering, community detection, degree-corrected stochastic block model, K-means, regularization.

1. Introduction

Advances in modern technology have facilitated the collection of network data which emerge in many fields including biology, bioinformatics, physics, economics, sociology and so forth. Therefore, developing effective analytic tools for network data has become a focal area in statistics research over the past decade. Network data often have natural communities which are groups of interacting objects (i.e., nodes); pairs of nodes in the same group tend to interact more often than pairs belonging to different groups. For example, in social networks, communities can be groups of people who belong to the same club, be of the same profession, or attend the same school; in protein-protein interaction networks, communities are regulatory modules of interacting proteins. In many cases, however, the underlying structure of network data is not directly observable. In such cases, we need to infer the latent community structure of nodes from knowledge of their interaction patterns.

The stochastic block model (SBM) proposed by **?** is a random graph model tailored for clustering nodes, and it is commonly used for recovering the community structure in network data. SBM has one limitation: it assumes that all nodes in the same community are stochastically equivalent (i.e., they have the same expected degrees). To overcome this limitation, **?** propose the degree-corrected stochastic block model (DCSBM) which allows for degree heterogeneity within communities. In the literature, various methods have been proposed for the estimation of SBM and DCSBM. They include but are not limited to modularity maximization (**?**), likelihood-based methods (**????**), the method of moments (**?**), spectral clustering (**?????????**), and spectral embedding (**??**). In most, if not all, works, theoretical properties such as consistency and asymptotic distributions are built based on the assumption that the true number of communities K_0 is known.

In practice, prior information of the number of communities is often unavailable. Accurately estimating K_0 from the network data is of crucial importance, as the following community detection procedure relies upon it. Determining the number of communities can be regarded as a model selection problem. A natural approach to the problem is to consider the popular model selection methods such as cross-validation (CV) or likelihood-based methods. However, tailoring those methods for SBMs or DCSBMs and establishing the theoretical support are challenging, as network data are complex in nature.

A few methods have been developed to estimate K_0. Among them, the eigenvalue-based methods have been widely applied; see **?**, **?**, **?** and **?** for the hypothesis testing methods on eigenvalues. These methods can be computationally fast, but they only use partial information from the data – the eigenvalues. Empirically, the good behavior of eigenvalues often requires a very large sample size. In order to make use of all the information from the data, we need to estimate the graph model (SBM or DCSBM). To this end, spectral clustering is considered as a quick and effective way, and it has been proven to have reliable theoretical basis (**???????**). Based on the spectral clustering method for estimating the graph model, **?** and **?** propose network cross-validation (NCV) and edge cross-validation (ECV), respectively, for selecting the number of communities. In particular, **?** show that the NCV method guarantees against under-selection in SBMs, but it does not rule out possible over-selection. Although they have a discussion on the estimation of DCSBMs, they do not study the theoretical property of the NCV estimator of the number of communities (K) in DCSBMs. **?** propose an ECV method for choosing between SBMs and DCSBMs along with selecting K for each model, but the consistency of ECV is not established. Moreover, both methods can be computationally intensive when the number of folds is large; they can lead to unstable results when the number of folds or the number of random sample splittings (or repetitions in the ECV case) is small. Another appealing method for model selection is the likelihood-based approach (**??**). It uses a BIC-type penalty, so that it avoids iterations or random sample splittings. However, for either SBMs or DCSBMs, optimizing the likelihood function which involves summing over all possible community memberships is computationally intractable for even moderate sample sizes. As a result, **?** use a variational EM algorithm to approximate the likelihood, and it may converge to a local optimum (**?**).

In this article, we propose a new method by taking advantage of both spectral clustering and likelihood principle. The method is devised for DCSBM, but can be naturally applied to SBM as it is a special case of DCSBM. To determine the number of communities K, we propose a pseudo likelihood ratio (pseudo-LR) to compare the goodness-of-fit of two DCSBMs estimated by using K and $K+1$, respectively, as the number of communities. For estimation, directly using spectral clustering can be an appealing choice as it is computationally fast. However, when $K > K_0$, it remains unclear about theoretical properties for the resulting estimators of the DCSBM obtained through the standard spectral clustering approach. This hinders the use of goodness-of-fit methods for model selection by spectral clustering for estimation. To overcome the difficulty, we estimate the DCSBM with K communities by spectral clustering; based on this estimate, we propose a binary segmentation method for estimating the DCSBM with $K + 1$ communities. This approach guarantees consistency of the estimator for the model with $K + 1$ communities when the estimator for the model with K communities is consistent. The binary segmentation technique has been used in the seminal work **?** for change-point detection and in recent work **?** for latent group recovery. Our idea of adapting this method to estimate DCSBM has not been considered by others. Based on the proposed estimation approach, we show that the pseudo-LR has a sound theoretical basis, and the resulting estimator of the number of communities is consistent.

It is worth noting that for establishing the consistency of estimating K_0, we only require the average degree to grow with the number of nodes n at a rate no slower than $\log(n)$, whereas the BIC-type methods considered in **?** and **?** require it to be faster than $n^{1/2}\log(n)$ and proportion to n, respectively, in DCSBMs. That is, these approaches need a much denser network than our method for good finite sample performance. As pointed out by **?**, Section 2.5, their approach needs a very stringent condition on the average degree, because the slow convergence rate of the estimate of the node degree variation passes on to the likelihood ratio. On the contrary, it is not carried on to our pseudo-LR because of the mutual cancellation of the slow-convergence parts. As a result, this allows us to relax the strong restriction on the average degree in theory. Both **?** and **?** only require the growth rate of the average degree to be no slower than $\log(n)$, which is the same rate as required by our method. However, theoretical properties are not available for the NCV and ECV estimators of K in DCSBMs. In contrast, we develop thorough theoretical results including the consistency of our proposed pseudo-LR method.

The rest of the paper is organized as follows. We describe the estimation procedure in Section 2. We establish the consistency of our estimators of the number of communities under DCSBMs in Section 3. Section 4 compares the performance of our method with various existing methods in different simulated networks. Section 5 illustrates the proposed method using several real data examples. Section 6 concludes. The proofs of all results are relegated to the appendix.

Notation. Throughout the paper, we write $[M]_{ij}$ as the (i, j)-th entry of matrix M. Without confusion, we sometimes simplify $[M]_{ij}$ as M_{ij}. In addition, we write $[M]_i$ as the i-th row of M. $\|M\|$ and $\|M\|_F$ denote the spectral norm and Frobenius norm of M, respectively. Note that $\|M\| = \|M\|_F$ when M is a vector. We use $\mathbf{1}\{\cdot\}$ to denote the indicator function which takes value 1 when $\cdot$ holds and 0 otherwise. All vectors without transpose are understood as column vectors. For a vector $\boldsymbol{a} = (a_1, ..., a_n)^\top$, let diag$(\boldsymbol{a})$ be the diagonal matrix whose diagonal is $\boldsymbol{a}$, and let $||\boldsymbol{a}|| = (\sum_i a_i^2)^{1/2}$ be its L$_2$ norm. Let ι_n, $\#\mathcal{S}$, and $[n]$ be the n-dimensional vector of ones, the cardinality of set $\mathcal{S}$, and the integer sequence $\{1, 2, \cdots, n\}$, respectively. C, c, and c' denote arbitrary positive constants that are independent of n, but may not be the same in different contexts.

2. Methodology

2.1 Degree-corrected SBM

Let $A \in \{0, 1\}^{n\times n}$ be the adjacency matrix. By convention, we do not allow self-connection, i.e., $A_{ii} = 0$. The network is generated by a degree-corrected stochastic block model with K_0 true communities. The communities, which represent a partition of the n nodes, are assumed to be fixed beforehand. Denote $Z_{K_0} = \{[Z_{K_0}]_{ik}\}$ as the $n \times K_0$ binary matrix providing the true cluster memberships of each node, i.e., $[Z_{K_0}]_{ik} = 1$ if node i is in $\mathcal{C}_{k,K_0}$ and $[Z_{K_0}]_{ik} = 0$ otherwise, where $\mathcal{C}_{1,K_0}, \ldots, \mathcal{C}_{K_0,K_0}$ are denoted as the communities identified by Z_{K_0}. For $k = 1, \cdots, K_0$, let $n_{k,K_0} = \#\mathcal{C}_{k,K_0}$, the number of nodes in $\mathcal{C}_{k,K_0}$. Given the K_0 communities, the edges between nodes i and j are chosen independently with probability depending on the communities that nodes i and j belong to. In particular, for nodes i and j belonging to clusters $\mathcal{C}_{k,K_0}$ and $\mathcal{C}_{l,K_0}$, respectively, the probability of edge between i and j is given by

$$P_{ij} = E(A_{ij}) = \theta_i \theta_j B_{kl,K_0},$$

where the block probability matrix $B_{K_0} = \{B_{kl,K_0}\}$, $k, l = 1, \ldots, K_0$, is a symmetric matrix with each entry between $(0, 1]$. The $n \times n$ edge probability matrix $P = \{P_{ij}\}$ represents the population counterpart of the adjacency matrix A. Let $\Theta = \text{diag}(\theta_1, \ldots, \theta_n)$. Then we have

$$P = E(A) = \Theta Z_{K_0} B_{K_0} Z_{K_0}^T \Theta^T.$$

Note that Θ and B_{K_0} are only identifiable up to scale. Following the lead of **?**, Theorem 3.3, we adopt the following normalization rule:

$$\sum_{i \in \mathcal{C}_{k,K_0}} \theta_i = n_{k,K_0}, \quad k = 1, \ldots, K_0. \tag{1}$$

Apparently, the DCSBM becomes the standard SBM when $\theta_i = 1$ for each $i = 1, ..., n$.

2.2 Estimation of the number of communities

Our procedure of estimating K_0 requires to obtain two estimated membership matrices $(\hat{Z}_K, \hat{Z}_{K+1}^b)$ based on K and $K + 1$ communities, respectively.[1] To this end, we estimate $\hat{Z}_K$ and $\hat{Z}_{K+1}^b$ via spectral clustering of the first K eigenvectors of the graph Laplacian and a binary segmentation technique, respectively. Section 2.3 provides more details. Denote $\hat{P}_{ij}(Z)$ as the estimator of P_{ij} for a given membership matrix Z. We compute $\hat{P}_{ij}(\hat{Z}_{K+1}^b)$ and $\hat{P}_{ij}(\hat{Z}_K)$ by the sample-frequency-type estimators and propose a pseudo-LR $L_n(\hat{Z}_{K+1}^b, \hat{Z}_K)$ defined in (2) to measure the deviance of goodness-of-fit of DCSBMs estimated with K and $K+1$ communities, respectively. The estimators of $\hat{P}_{ij}(\hat{Z}_{K+1}^b)$ and $\hat{P}_{ij}(\hat{Z}_K)$ are given in Appendix A. Lastly, we obtain the estimator of the true number of communities based on the change of the pseudo-LR. Let $K_{\max}$ denote the maximum number of communities such that $K_{\max} \geq K_0$. The pseudo-code is described in Algorithm 1.

[1] The superscript b in $\hat{Z}_{K+1}^b$ denotes that it is estimated by a binary segmentation from $\hat{Z}_K$.

Algorithm 1: Estimation of the number of communities

```
input  : adjacency matrix $A$, tuning parameters $c_\eta$ and $h_n$
output: $\hat{K}_1$ and $\hat{K}_2$
for $K \leftarrow 1$ to $K_{\max}$ do
    obtain $\hat{Z}_K$ and $\hat{Z}^b_{K+1}$ via spectral clustering and binary segmentation, respectively;
    compute $\hat{P}_{ij}(\hat{Z}_K)$ and $\hat{P}_{ij}(\hat{Z}^b_{K+1})$;
    compute
        $L_n(\hat{Z}^b_{K+1}, \hat{Z}_K) = \frac{1}{2} \sum_{i \neq j} \left( \frac{\hat{P}_{ij}(\hat{Z}^b_{K+1})}{\hat{P}_{ij}(\hat{Z}_K)} - 1 \right)^2$ [a]    (2)
    compute $R(K)$ as
        $R(K) = \begin{cases} \frac{L_n(\hat{Z}^b_{K+1}, \hat{Z}_K)}{\eta_n} & K = 1 \\ \frac{L_n(\hat{Z}^b_{K+1}, \hat{Z}_K)}{L_n(\hat{Z}^b_K, \hat{Z}_{K-1})} & K \geq 2, \end{cases}$    (3)
    where $\eta_n = c_\eta n^2$.
end
obtain $\hat{K}_1$ and $\hat{K}_2$ as
    $\hat{K}_1 = \arg\min_{1 \leq K \leq K_{\max}} R(K),$
  and
    $\hat{K}_2 = \min(\hat{K}_1, \tilde{K}_2),$
  where $\tilde{K}_2 = \min\{K \in \{1, \cdots, K_{\max}\}, R(K) \leq h_n\}$ if $\min_{1 \leq K \leq K_{\max}} R(K) \leq h_n$
  and $\tilde{K}_2 = K_{\max}$ otherwise.
```

[a] If $\hat{P}_{ij}(\hat{Z}_K)$ in the denominator is zero, one can replace it by a constant that is sufficiently small, say, 2^{-52}.

To understand our algorithm of estimating K_0, we focus on the case where $K_0 \geq 2$. If we know that $K_0 \geq 2$ for sure, we can redefine $\hat{K}_1 = \arg\min_{2 \leq K \leq K_{\max}} R(K)$. By Theorems 5 and 6 in Section 3.3, we have

$$L_n(\hat{Z}^b_K, \hat{Z}_{K-1}) \asymp n^2 \text{ for } 2 \leq K \leq K_0 \text{ and } L_n(\hat{Z}^b_{K_0+1}, \hat{Z}_{K_0}) \leq O_{a.s.}(n\rho_n^{-1}),$$

where $a_n \asymp b_n$ means that $P(c \leq a_n/b_n \leq C) \to 1$ as $n \to \infty$ for some positive constants c and C, $a.s.$ denotes almost surely, and the parameter ρ_n characterizes the sparsity of the network such that $n\rho_n/\log(n)$ is sufficiently large (see Assumption 4 in Section 3.2). This result directly implies that

$$R\left(K\right) \asymp 1 \text{ for } 2 \leq K < K_0 \text{ and } R\left(K_0\right) = o_p\left(1\right).$$

The above results indicate that for $K = K_0$, $R\left(K\right)$ is very small and close to zero, but for $K < K_0$, $R\left(K\right)$ is relatively large. It is worth noting that for $K > K_0$, it is possible that $R\left(K\right)$ is also small. As a result, the minimizer of $R(K)$ is only guaranteed to satisfy $\hat{K}_1 \geq K_0$ with probability approaching 1 (w.p.a.1) as $n \to \infty$. Such a result is similar to that in **?** who show that NCV do not underestimate the number of communities w.p.a.1 as $n \to \infty$. Based on our theory, we expect to observe a gap of the values of $R\left(K\right)$ at $K = K_0$, so we introduce $\tilde{K}_2$ which is the first K such that

$R(K)$ is less than h_n, where $h_n \to 0$ and $n\rho_n h_n \to \infty$. Then we have $\tilde{K}_2 = K_0$ w.p.a.1 as $n \to \infty$. For better numerical performance, we make use of both $\hat{K}_1$ and $\tilde{K}_2$ by letting $\hat{K}_2 = \min(\hat{K}_1, \tilde{K}_2)$, and thus it satisfies $P(\hat{K}_2 = K_0) \to 1$ as $n \to \infty$, i.e., $\hat{K}_2$ consistently estimates the number of communities in large samples. In our algorithm, two tuning parameters c_η and h_n are involved. Among them, c_η is only needed to deal with the case $K = 1$ in which the pseudo-LR cannot be defined. If we are sure that $K_0 \geq 2$, i.e., there are more than one communities, we can obtain the estimate $\hat{K}_1$ by searching over $K \in [2, K_{\max}]$. Alternatively, one can separately test $K_0 = 1$ using other methods and then use our methods to select K for $K \geq 2$. [2] In both cases, one can avoid the use of c_η. Theoretically, c_η only needs to satisfy $c_\eta \in (0, \infty)$. Practically, We choose a value for c_η given in Section 4.3 that works well in our numerical analysis. For the choice of h_n, we have a detailed discussion given after Theorem 6 in Section 3.3.

Our pseudo-LR defined in (2) has a connection with the original likelihood ratio (LR) statistic:

$$\sum_{i \neq j} A_{ij} \log \left(\frac{\hat{P}_{ij}(\hat{Z}^b_{K+1})}{\hat{P}_{ij}(\hat{Z}_K)} \right) + (1 - A_{ij}) \log \left(\frac{1 - \hat{P}_{ij}(\hat{Z}^b_{K+1})}{1 - \hat{P}_{ij}(\hat{Z}_K)} \right).$$

In the semi-dense networks where P_{ij} decays to zero, $\log \left(\frac{1-\hat{P}_{ij}(\hat{Z}^b_{K+1})}{1-\hat{P}_{ij}(\hat{Z}_K)} \right)$ is approximately given by $\log(1) = 0$ so that the second term in the above display is asymptotically negligible. As a result, the major contribution to the LR statistic is attributable to $\log \left(\frac{\hat{P}_{ij}(Z_K)}{\hat{P}_{ij}(Z_{K-1})} \right)$. This motivates us to construct our pseudo-LR defined in (2). Moreover, the estimate of K constructed directly from the original LR requires a more restrictive assumption on the average degree than our estimate obtained from the pseudo-LR (c.f., (**?**)), as the slow-convergent estimates of the node degree parameters that are involved in the original LR statistic are mutually cancelled out by our pseudo-LR.

The pseudo-LR statistic in (2) is also connected to the traditional χ^2-test for goodness of fit (**??**) where the test statistic can be written as $\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$, where o_i denotes the observed frequency of the ith category and e_i denotes the expected frequency of the ith category under the null hypothesis. Accordingly, a pseudo χ^2-type test statistic can be defined as follows

$$L_n^0 \left(\hat{Z}^b_{K+1}, \hat{Z}_K \right) = \frac{1}{2} \sum_{i \neq j} \frac{\left[\hat{P}_{ij}(\hat{Z}^b_{K+1}) - \hat{P}_{ij}(\hat{Z}_K) \right]^2}{\hat{P}_{ij}(\hat{Z}_K)} = \frac{1}{2} \sum_{i \neq j} \left[\frac{\hat{P}_{ij}(\hat{Z}^b_{K+1})}{\hat{P}_{ij}(\hat{Z}_K)} - 1 \right]^2 \hat{P}_{ij}(\hat{Z}_K),$$

where $\hat{P}_{ij}(\hat{Z}^b_{K+1})$ and $\hat{P}_{ij}(\hat{Z}_K)$ play the roles of o_i and e_i, respectively. Nevertheless, due to the fact that $\hat{P}_{ij}(\hat{Z}_K)$ can shrink to zero in sparse networks, the investigation of the asymptotic behavior of $L_n^0 \left(\hat{Z}^b_{K+1}, \hat{Z}_K \right)$ is not as convenient as that of the pseudo-LR statistic in (2), which can be regarded as a weighted version of $L_n^0 \left(\hat{Z}^b_{K+1}, \hat{Z}_K \right)$.

2.3 Estimation of the memberships

The proposed pseudo-LR given in (2) depends on $(\hat{Z}_K, \hat{Z}^b_{K+1})$ which are obtained through spectral clustering and binary segmentation, respectively. In the following, we describe the algorithm in

[2] **?** propose one method based on the limiting distribution of the principal eigenvalue of the suitably centered and scaled adjacency matrix generated from a SBM. To extend such result to DCSBM is an interesting topic for future research.

detail. Let $\hat{d}_i = \sum_{j=1}^n A_{ij}$ denote the degree of node i, $D = \text{diag}(\hat{d}_1, \ldots, \hat{d}_n)$. We regularize the degree for each node as $\hat{d}_i^\tau = \hat{d}_i + \tau$ where τ is a regularization parameter. Let $D_\tau = \text{diag}(\hat{d}_1 + \tau, \ldots, \hat{d}_n + \tau)$. The regularized sample graph Laplacian is

$$L_\tau = D_\tau^{-1/2} A D_\tau^{-1/2}.$$

We regularize the sample degree matrix D to improve the finite sample performance of spectral clustering. The same regularization strategy is considered by **?**, **?** and **?**. The corresponding theoretical property is established in Section 3.

Denote the spectral decomposition of L_τ as

$$L_\tau = \widehat{U}_n \widehat{\Sigma}_n \widehat{U}_n^T,$$

where $\widehat{\Sigma}_n = \text{diag}(\hat{\sigma}_{1n}, \ldots, \hat{\sigma}_{nn})$ with $|\hat{\sigma}_{1n}| \geq |\hat{\sigma}_{2n}| \geq \cdots \geq |\hat{\sigma}_{nn}| \geq 0$, and $\widehat{U}_n$ is the corresponding eigenvectors such that $\widehat{U}_n^T \widehat{U}_n = I_n$. For each $K = 1, \cdots, K_{\max}$, let

$$\hat{\nu}_{iK} = \frac{\hat{u}_i(K)}{||\hat{u}_i(K)||}, \tag{4}$$

where $\hat{u}_i^T$ is the i-th row of $\widehat{U}_n$ and $\hat{u}_i^T(K)$ collects the first K elements of $\hat{u}_i^T$. We estimate the pair of community memberships $(\hat{Z}_K, \hat{Z}_{K+1}^b)$ by the following algorithm.

Algorithm 2: Estimation of the memberships

```
input  : $\{\hat{\nu}_{iK}, \hat{\nu}_{iK+1}\}_{i=1}^n$ and $K$
output: $\hat{Z}_K$ and $\hat{Z}_{K+1}^b$
divide $\{\hat{\nu}_{iK}\}_{i=1}^n$ into $K$ groups by the k-means algorithm with $K$ centroids. Denote the
  membership matrix as $\hat{Z}_K$ with the corresponding communities $\{\widehat{\mathcal{C}}_{k,K}\}_{k=1}^K$;
for $k \leftarrow 1$ to $K$ do
    divide $\widehat{\mathcal{C}}_{k,K}$ into two subgroups by applying the k-means algorithm on
      $\{\hat{\nu}_{iK+1}\}_{i \in \widehat{\mathcal{C}}_{k,K}}$. Denote the two subgroups as $\widehat{\mathcal{C}}_{k,K}(1)$ and $\widehat{\mathcal{C}}_{k,K}(2)$;
    compute
        $\widehat{Q}_K(k) = \frac{\widehat{\Phi}(\widehat{\mathcal{C}}_{k,K}) - \widehat{\Phi}(\widehat{\mathcal{C}}_{k,K}(1)) - \widehat{\Phi}(\widehat{\mathcal{C}}_{k,K}(2))}{\#\widehat{\mathcal{C}}_{k,K}},$        (5)
      where for an arbitrary index set $C$, $\widehat{\Phi}(C) = \sum_{i \in \mathcal{C}} ||\hat{\nu}_{iK+1} - \frac{\sum_{i \in \mathcal{C}} \hat{\nu}_{iK+1}}{\#\mathcal{C}}||^2$;
end
choose $\hat{k} = \arg\max_{1 \leq k \leq K} \hat{Q}_K(k)$ and denote
    $\{\widehat{\mathcal{C}}_{k,K+1}^b\}_{k=1}^{K+1} = \{\{\widehat{\mathcal{C}}_{k,K}\}_{k<\hat{k}}, \widehat{\mathcal{C}}_{\hat{k},K}(1), \{\widehat{\mathcal{C}}_{k,K}\}_{k>\hat{k}}, \widehat{\mathcal{C}}_{\hat{k},K}(2)\}$
  as the new groups for $K+1$. The corresponding membership matrix is denoted as
  $\hat{Z}_{K+1}^b$.
```

Algorithm 2 applies the standard spectral clustering approach to obtain $\hat{Z}_K$ and a binary segmentation method to obtain $\hat{Z}_{K+1}^b$. This procedure is computationally fast. Moreover, the algorithm leads to $\widehat{\mathcal{C}}^b_{k,K+1} = \widehat{\mathcal{C}}_{k,K}$ for $k \neq \hat{k}$ and $\widehat{\mathcal{C}}^b_{\hat{k},K+1} \cup \widehat{\mathcal{C}}^b_{K+1,K+1} = \widehat{\mathcal{C}}_{\hat{k},K}$, which ensures that the parameter estimators $\hat{P}_{ij}(\hat{Z}_K)$ and $\hat{P}_{ij}(\hat{Z}^b_{K+1})$ in the DCSBM are consistent when $K = K_0$.

It is worth mentioning that the binary segmentation step is crucial for our theoretical derivation. To use our pseudo LR statistic in (2), we need to obtain the estimated group memberships at K and $K+1$, respectively. When $K = K_0$, the estimates of the memberships for $(K+1)$ obtained from the spectral clustering no longer have a theoretical guarantee, as the $(K+1)$-th largest eigenvalue of the population Graph Laplacian is zero, and the corresponding $(K+1)$-th column of the eigenvector matrix is not uniquely identified. On the other hand, the binary segmentation step obtains the $(K+1)$-th group via dividing one of the K groups into two groups. This nested structure ensures estimation consistency of the model with $(K+1)$ groups as long as the estimator of the model with K communities is consistent.

3. Theory

3.1 Identification

The population counterpart of L_τ is

$$\mathcal{L}_\tau = \mathcal{D}_\tau^{-1/2} P \mathcal{D}_\tau^{-1/2},$$

where $\mathcal{D}_\tau = \mathcal{D} + \tau I_n$ and $\mathcal{D} =$ diag$(d_1, \ldots, d_n)$ with $d_i = \sum_{j=1}^n P_{ij}$. Let $\pi_{kn} = n_{k,K_0}/n$ and $\Pi_n =$ diag$(\pi_{1n}, \cdots, \pi_{K_0 n})$.

Assumption 1 *Let $H_{K_0} = \rho_n^{-1} B_{K_0} = [H_{kl,K_0}]$ for some $\rho_n > 0$, $W_k = \sum_{l=1}^{K_0} H_{kl,K_0}\pi_{ln}$, $\mathcal{D}_H =$ diag$(W_1, \cdots, W_{K_0})$, and $H_{0,K_0} = \mathcal{D}_H^{-1/2} H_{K_0} \mathcal{D}_H^{-1/2}$. Then,*

(1) H_{K_0} is not varying with n,

(2) as $n \to \infty$, $H_{0,K_0} \to H^*_{0,K_0}$ where H^*_{0,K_0} has full rank K_0 and K_0 is fixed,

(3) all elements of H^*_{0,K_0} are positive,

(4) there exist two constants $\underline{\theta}$ and $\overline{\theta}$ such that $0 < \underline{\theta} \leq \inf_i \theta_i \leq \sup_i \theta_i \leq \overline{\theta}$.

Several remarks are in order. First, Assumption 1 implies that the average node degree is of order $n\rho_n$. The network can be semi-dense if $\rho_n \to 0$ but $n\rho_n \to \infty$. Second, Assumption 1(1) is just for notational simplicity. All our results still hold if H_{K_0} depends on n and converges to some limit. Third, Assumption 1(2) ensures that the DCSBM has K_0 communities. To see this, note that Assumption 1(2) implies both H_{K_0} and B_{K_0} have full rank. Suppose there exist $\{\tilde{\theta}_i\}_{i=1}^n$, $\tilde{\Theta} =$ diag$(\tilde{\theta}_1, \cdots, \tilde{\theta}_n)$, $\tilde{Z}_{K_0'}$, and $\tilde{B}_{K_0'}$ such that $\tilde{B}_{K_0'}$ is a full rank $K_0' \times K_0'$ matrix and

$$\Theta Z_{K_0} B_{K_0} Z_{K_0}^T \Theta^T = P = \tilde{\Theta} \tilde{Z}_{K_0'} \tilde{B}_{K_0'} \tilde{Z}_{K_0'}^T \tilde{\Theta}^T.$$

Further suppose that the membership matrix $\tilde{Z}_{K_0'}$ is non-degenerate in the sense that each community identified by $\tilde{Z}_{K_0'}$ is nonempty, which implies that $\tilde{Z}_{K_0'}$ has full column rank. Then, the full rank condition of B_{K_0} and $\tilde{B}_{K_0'}$ implies that

$$\begin{aligned} K_0 = \text{rank}(B_{K_0}) =& \text{rank}(\Theta Z_{K_0} B_{K_0} Z_{K_0}^T \Theta^T) \\ =& \text{rank}(P) \\ =& \text{rank}(\tilde{\Theta} \tilde{Z}_{K_0'} \tilde{B}_{K_0'} \tilde{Z}_{K_0'}^T \tilde{\Theta}^T) = \text{ rank}(\tilde{B}_{K_0'}) = K_0'. \end{aligned}$$

That is, the number of communities is identified. It is possible that some SBMs can be rewritten as DCSBMs with specific degree corrections. For example, a SBM with $K_0 = 2$ and the block probability

$$B = \begin{pmatrix} 4/9 & 2/9 \\ 2/9 & 1/9 \end{pmatrix}$$

can be viewed as a DCSBM with $K_0 = 1$, $B = 4/9$, and

$$\theta_i = \begin{cases} 1 & \text{if node i belongs group 1} \\ 1/2 & \text{if node i belongs group 2} \end{cases}.$$

In this case, the full rank condition rules out the SBM model with $K_0 = 2$ and the model of DCSBM with $K_0 = 1$ is the one considered in the paper. Fourth, from the perspective of real data applications, the full-rank condition on B_{K_0} is reasonable. In networks, communities are usually groups of nodes that have a higher probability of being connected to each other within the same group than to members of other groups. This directly implies the full rank condition of B_{K_0} if $K_0 = 2$. In general, by the Gershgorin circle theorem, for each row, if the sum of off-diagonal elements is strictly less than the diagonal element, i.e., for $k = 1, \cdots, K_0$

$$\sum_{l=1,\cdots,K_0,\, l \neq k} B_{kl,K_0} < B_{kk,K_0},$$

then B has full rank. Such condition is just a sufficient condition for our full rank requirement. For estimating the SBMs, the semi-definite programming method can also be used. It needs the strong assortativity condition (**?**) given as

$$\min_{k=1,\cdots,K_0} B_{kk,K_0} > \max_{k,l=1,\cdots,K_0,\, k \neq l} B_{kl,K_0}.$$

In general, the strong assortativity and Assumption 1(2) do not nest within each other. For example, the following matrix has full rank but violates the strong assortativity:

$$\begin{pmatrix} 0.8 & 0.4 & 0.1 \\ 0.4 & 0.5 & 0.05 \\ 0.1 & 0.05 & 0.2 \end{pmatrix}.$$

Which assumption is more plausible depends on the empirical data at hand. In the three real data examples considered in Section 5 and Appendix C, the full rank condition holds for all of them, but the strong assortativity does not hold for the political books network. Fifth, from the theoretical perspective, the full-rank condition (i.e., the K_0-th largest absolute eigenvalue of the $\mathcal{L}_\tau$ is bounded away from zero) is a common assumption in the literature. See, for example, **?**, **?**, and **?**. It is fundamental for the spectral clustering method. If it does not hold, i.e., the K_0-th eigenvalue of the population graph Laplacian is exactly zero, then the corresponding population eigenvector is not uniquely defined. Sixth, Assumption 1(3) is a technical condition which is sufficient for $\hat{\nu}_{iK}$ in (4) to be well-defined, as shown in Lemma 7 in Appendix E. Last, for simplicity, we restrict θ_i to be bounded between zero and infinity. This assumption can be relaxed at the cost of more complicated notations.

Next, let $\Theta_\tau = \text{diag}(\theta_1^\tau, \dots, \theta_n^\tau)$, where $\theta_i^\tau = \theta_i d_i/(d_i + \tau)$ for $i = 1, \dots, n$, $n_{k,K_0}^\tau = \sum_{i \in \mathcal{C}_{k,K_0}} \theta_i^\tau$, and $\Pi_n^\tau = \text{diag}(n_{1,K_0}^\tau/n, \cdots, n_{K_0,K_0}^\tau/n)$.

Assumption 2 *Suppose*

(1) there exist $\{\pi_{k\infty}\}_{k=1}^{K_0}$ and $\{\pi'_{k\infty}\}_{k=1}^{K_0}$ that are bounded between zero and infinity such that

$$\Pi_n \to \Pi_\infty = \textit{diag}(\pi_{1\infty}, \dots, \pi_{K_0\infty}) \quad \textit{and} \quad \Pi_n^\tau \to \Pi'_\infty = \textit{diag}\ (\pi'_{1\infty}, \dots, \pi'_{K_0\infty}),$$

(2) $(\Pi'_\infty)^{1/2} H_{0,K_0}^* (\Pi'_\infty)^{1/2}$ has K_0 distinct eigenvalues.

The second convergence in Assumption 2(1) can be easily satisfied by choosing τ to be the average degree $(\bar{d})$ in the network. Let $|\lambda_1| \geq \cdots \geq |\lambda_{K_0}|$ be the eigenvalues of $(\Pi'_\infty)^{1/2} H_{0,K_0}^* (\Pi'_\infty)^{1/2}$ and

$$\text{eigsp}((\Pi'_\infty)^{1/2} H_{0,K_0}^* (\Pi'_\infty)^{1/2}) = \min_{k=1,\cdots,K_0-1} |\lambda_{k+1} - \lambda_k|$$

be the gap between adjacent eigenvalues of $(\Pi'_\infty)^{1/2} H_{0,K_0}^* (\Pi'_\infty)^{1/2}$, as defined in **?**. Then, Assumption 2(2) requires that

$$\text{eigsp}((\Pi'_\infty)^{1/2} H_{0,K_0}^* (\Pi'_\infty)^{1/2}) \geq C > 0$$

for some constant C. The same condition is assumed in **?**.[3] Assumption 2(2) is mild from a practical point of view. If we denote H_{0,K_0}^* as $vec(H_{0,K_0}^*) \in \Re^{K_0^2}$ such that H_{0,K_0}^* is symmetric and full rank, then Assumption 2(2) is only violated for a set in $\Re^{K_0^2}$ with zero Lebesgue measure. Theoretically, as K_0 is not known a priori, we need to apply spectral clustering to the first K eigenvectors of the graph Laplacian for $K = 1, \cdots, K_0$. Therefore, at the population level, we require that the eigenspace generated by the first K eigenvectors is identified for all $K = 1, \cdots, K_0$, which is equivalent to Assumption 2(2).

Consider the spectral decomposition of $\mathcal{L}_\tau$,

$$\mathcal{L}_\tau = U_{1n} \Sigma_{1n} U_{1n}^T,$$

where $\Sigma_{1n} = \text{diag}(\sigma_{1n}, \dots, \sigma_{K_0 n})$ is a $K_0 \times K_0$ matrix that contains the eigenvalues of $\mathcal{L}_\tau$ such that $|\sigma_{1n}| \geq |\sigma_{2n}| \geq \cdots \geq |\sigma_{K_0 n}| > 0$ and $U_{1n}^T U_{1n} = I_{K_0}$.

Theorem 1 *Suppose Assumptions 1 and 2 hold. Let u_i^T and $u_i(K)$ be the i-th row of U_{1n} and the top K elements of u_i, respectively.*

(1) If $[Z_{K_0}]_i = [Z_{K_0}]_j$, then $\|\frac{u_i}{\|u_i\|} - \frac{u_j}{\|u_j\|}\| = 0$; if $[Z_{K_0}]_i \neq [Z_{K_0}]_j$, then $\|\frac{u_i}{\|u_i\|} - \frac{u_j}{\|u_j\|}\| = \sqrt{2}$.

[3] See **?**, Lemma 2.3.

(2) *There exist L_K distinct $K \times 1$ vectors, denoted as $(\bar{\nu}_{1,K}, \cdots, \bar{\nu}_{L_K,K})$, such that the nodes can be divided into L_K groups, denoted by $\{G_{l,K}\}_{l=1}^{L_K}$, $K \leq L_K \leq K_0$, for any $l = 1, \cdots, L_K$,*

$$\limsup_n \sup_{i,j \in G_{l,K}} \left\| \frac{u_i(K)}{||u_i(K)||} - \bar{\nu}_{l,K} \right\| = 0,$$

and for any $l \neq l'$ and some constant $c > 0$ independent of n,

$$\liminf_n \inf_{i \in G_{l,K}, j \in G_{l',K}} \left\| \frac{u_i(K)}{||u_i(K)||} - \bar{\nu}_{l,K} \right\| \geq c.$$

Several remarks are in order. First, Theorem 1(1) has already been established in the literature. See **?** and **?**. It implies that the eigenvectors of the graph Laplacian contain information about the group structure. Second, Theorem 1(2) implies that the first K columns of eigenvectors after row normalization still contain information for at least K communities, when $K \leq K_0$. In particular, when $K = K_0$, $L_{K_0} = K_0$ and Theorem 1(1) implies that Theorem 1(2) holds with the true communities, i.e., $\{G_{l,L_{K_0}}\}_{l=1}^{L_{K_0}} = \{\mathcal{C}_{k,K_0}\}_{k=1}^{K_0}$. Therefore, $\{G_{l,K}\}_{l=1}^{L_K}$ can be viewed as the true communities identified by the first K columns of eigenvectors. Third, Lemma 7 in Appendix E implies that $||u_i(K)||$ is bounded away from zero for $K = 1, \cdots, K_0$, which guarantees that $\frac{u_i(K)}{||u_i(K)||}$ is well defined. This result is similar to **?**, Lemma 2.5.

3.2 Properties of the estimated memberships

In the following, we aim to show that, under certain conditions, if $K \leq K_0$, then $\hat{Z}_K = Z_K$ and $\hat{Z}_K^b = Z_K^b$ almost surely (a.s.) for some deterministic membership matrices Z_K and Z_K^b. We denote the communities identified by Z_K and Z_K^b as $\{\mathcal{C}_{k,K}\}_{k=1}^K$ and $\{\mathcal{C}_{k,K}^b\}_{k=1}^K$, respectively. Note that L_K is not necessarily equal to K. This implies that neither $\{\mathcal{C}_{k,K}\}_{k=1}^K$ nor $\{\mathcal{C}_{k,K}^b\}_{k=1}^K$ is necessarily equal to the true communities $\{G_{l,K}\}_{l=1}^{L_K}$. We can view Z_K and Z_{K+1}^b as the pseudo true values of our estimation procedure described in Section 2.2. We slightly abuse the notation by calling Z_K evaluated at $K = K_0$ as the pseudo true membership matrix when $K = K_0$ while Z_{K_0} as the true membership matrix. Theorem 4 below shows that when $K = K_0$, the pseudo true values Z_K and Z_K^b are equal to the true membership matrix Z_{K_0}. Therefore, the notation is still consistent and we can just write Z_{K_0} as the (pseudo) true membership matrix for $K = K_0$.

Definition 2 *For $i \in G_{l,K}$ and $l = 1, ..., L_K$, $K = 2, \cdots, K_0$, let*

$$\nu_{iK} = \bar{\nu}_{l,K}.$$

Then, (Z_K, Z_{K+1}^b) is defined by applying Algorithm 2 to $\{\nu_{iK}\}_{i=1}^n$, $K = 1, \cdots, K_0 - 1$. When $K = 1$, we can trivially define $Z_1 = Z_1^b = [n] = \{1, 2, ..., n\}$.

Assumption 3 *Suppose that*

(1) *the above definitions of Z_K and Z_K^b are unique for $K = 1, \cdots, K_0$;*

(2) *there exist a positive constant c independent of n and $k^* = 1, \cdots, K$ such that $Q_K(k^*) - \max_{k \neq k^*} Q_K(k) \geq c$ for $K = 2, \cdots, K_0 - 1$, where $Q_K(\cdot)$ is similarly defined as $\hat{Q}_K(\cdot)$ in* (5) *with $\hat{\nu}_{iK+1}$ and $\{\widehat{\mathcal{C}}_{k,K}\}$ replaced by ν_{iK+1} and $\{\mathcal{C}_{k,K}\}$, respectively.*

Several remarks are in order. First, the communities identified by Z_{K+1}^b can be written as

$$\{\mathcal{C}_{k,K+1}^b\}_{k=1}^{K+1} = \{\mathcal{C}_{1,K}, \cdots, \mathcal{C}_{k^*-1,K}, \mathcal{C}_{k^*,K}(1), \mathcal{C}_{k^*,K}(2), \mathcal{C}_{k^*+1,K}, \cdots, \mathcal{C}_{K,K}\}.$$

Second, we provide more details on Z_K, Z_K^b, and $Q_K(\cdot)$ in Appendix A. Third, the uniqueness requirement is mild. If $L_K = K$, then obviously $\{\mathcal{C}_{k,K}\}_{k=1}^K = \{G_{l,K}\}_{l=1}^{L_K}$, which implies Z_K is uniquely defined. Fourth, we have $L_{K_0} = K_0$. Therefore, by definition, $\{\mathcal{C}_{k,K_0}\}_{k=1}^{K_0}$ defined by Z_{K_0} equal $\{G_{l,K_0}\}_{l=1}^{K_0}$, which are the true communities. Fifth, when $L_K = K$ and $L_{K+1} = K+1$ for $K \leq K_0 - 1$, by the pigeonhole principle, there only exists one $k \in \{1, \cdots, K\}$, denoted as $k^\dagger$ such that $\mathcal{C}_{k^\dagger,K} = G_{k^\dagger,K}$ contains two of $\{G_{l,K+1}\}_{l=1}^{K+1}$. Then by Theorem 1(2), there exists some constant $c > 0$ such that $Q_K(k^\dagger) \geq c$ and $Q_K(k) \to 0$ for $k \neq k^\dagger$. In this case, $k^* = k^\dagger$ and Assumption 3(2) holds. Sixth, Assumption 3 is similar to **?**, Assumption 2.1. It is used as a matter of notational convenience but not of necessity. Under Assumption 3, we will show that the pseudo-LR after re-centering is asymptotically normal. If Assumption 3 fails and (Z_K, Z_K^b) are not unique, it can be anticipated that the pseudo-LR after re-centering will be asymptotically mixture normal with weights depending on the probability of choosing one classification among all possibilities. Last, although Assumption 3 is used to characterize the limiting distribution of the re-centered pseudo-LR, it does not affect the rate of bias term in the under-fitting case. Because the bias term will dominate the centered term, we actually only need the rate of bias to show the validity of our selection procedure. Therefore, even if Assumption 3 fails, it is reasonable to expect that our procedure can still consistently select the true number of communities as established in Section 3.3.

Assumption 4 *Assume $\rho_n n/\log(n) \geq C_1$ for some constant $C_1 > 0$ sufficiently large and $\tau = O(n\rho_n)$.*

Recall that the degree of the network is of order $n\rho_n$. Assumption 4 requires the degree to diverge at a rate no slower than $\log(n)$, which is the most relaxed degree growth rate for exact community recovery when K is known. See **?** for an excellent survey on the recent development of estimation of SBMs and DCSBMs.[4] For determining the number of communities, **?** require the same condition on the degree for SBMs, but they do not provide any theory for DCSBMs. **?** establish the theories for DCSBMs but require that $n^{1/2}\rho_n/\log(n) \to \infty$, or equivalently, the degree diverges to infinity at a rate faster than $n^{1/2}\log(n)$. We require a weaker condition compared to **?**, mainly due to the fact that we use a pseudo instead of the true likelihood ratio. In DCSBMs, the rate of convergence for the estimator $\hat{\theta}_i$ of θ_i is much slower than that for the estimator of the block probability matrix. By using the ratio $\frac{\hat{P}_{ij}(\hat{Z}_{K+1}^b)}{\hat{P}_{ij}(\hat{Z}_K)}$ in the definition of pseudo-LR, the components of $\hat{\theta}_i$'s that cause the slower convergence rate in both the numerator and the denominator cancel each other out, so that the convergence rate of $\frac{\hat{P}_{ij}(\hat{Z}_{K+1}^b)}{\hat{P}_{ij}(\hat{Z}_K)}$ is unaffected. We recommend using regularization to improve the finite sample performance of spectral clustering. By Assumption 1, setting τ as the average degree $\bar{d}$ satisfies Assumption 4. In practice, $\bar{d}$ is unobserved and we replace it by the sample version, following the lead of **?**. In the proof of Theorem 5 in Appendix D, we show that the sample average degree is of the same order of magnitude as its population counterpart almost surely because

$$\sup_i \left| \frac{\hat{d}_i}{d_i} - 1 \right| \leq C\sqrt{\frac{\log(n)}{n\rho_n}}$$

[4] We thank a referee for this reference.

for some fixed constant $C > 0$. One can also use the data-driven method proposed by **?** to select the regularizer. Based on the simulation study in **?**, the performances of spectral clustering using sample average degree and data-driven regularizer are similar.

Definition 3 *Suppose there are two membership matrices Z_1 and Z_2 with corresponding communities $\{\mathcal{C}_k^j\}_{k=1}^{K_j}$, $j = 1, 2$, respectively. Then we say Z_1 is finer than Z_2 if for any $k_1 = 1, \cdots, K_1$, there exists $k_2 = 1, \cdots, K_2$ such that*

$$\mathcal{C}_{k_1}^1 \subset \mathcal{C}_{k_2}^2.$$

In this case, we write $Z_1 \succeq Z_2$.

Theorem 4 *If Assumptions 1–4 hold, then*

(1) for $K = 1, \cdots, K_0$,

$$\hat{Z}_K = Z_K \quad a.s. \quad and \quad Z_{K_0} \succeq Z_K,$$

(2) for $K = 1, \cdots, K_0 - 1$,

$$\hat{Z}_{K+1}^b = Z_{K+1}^b \quad a.s. \quad and \quad Z_{K_0} \succeq Z_{K+1}^b,$$

(3) after relabeling, we have $\widehat{\mathcal{C}}_{k,K+1}^b = \mathcal{C}_{k,K}$ for $k = 1, \cdots, K-1$ and $\mathcal{C}_{K,K} = \widehat{\mathcal{C}}_{K,K+1}^b \cup \widehat{\mathcal{C}}_{K+1,K+1}^b$, for $K = 1, \cdots, K_0$, a.s.

Theorem 4(1) and (2) show that $\hat{Z}_K$ and $\hat{Z}_K^b$ equal their pseudo true counterparts almost surely. This is the oracle property of estimating the community membership when we either under- or just-fit the model, i.e., $K \leq K_0$. On the other hand, it is very difficult, if not completely impossible, to show the similar oracle property for the over-fitting case, i.e., $K > K_0$. In particular, we are unable to uniquely define $Z_{K_0+1}^b$ and show that $\hat{Z}_{K_0+1}^b = Z_{K_0+1}^b$ *a.s.* As pointed out by **?**, even in the population level (i.e., the probability matrix is observed), "embedding a K-block model in a larger model can be achieved by appropriately splitting the labels Z and there are an exponential number of possible splits." However, Theorem 4(3) with $K = K_0$ shows that, for any $k = 1, \cdots, K_0 + 1$, there exists some k' such that $\widehat{\mathcal{C}}_{k,K_0+1}^b \subset \widehat{\mathcal{C}}_{k',K_0}$, which should be one of the true communities based on the oracle property. We can use this feature to handle the over-fitting case.

3.3 Properties of the pseudo-LR and the estimated number of communities

Without loss of generality, we assume that $\hat{Z}_K^b$ is obtained by splitting the last group in $\hat{Z}_{K-1}$ into the $(K-1)$-th and K-th groups in $\hat{Z}_K^b$. Further denote, for $k, l = 1, \cdots, K$ and $k \leq l$,

$$\Gamma_{kl,K}^{0b} = \sum_{s \in I(\mathcal{C}_{k,K}^b),\, t \in I(\mathcal{C}_{l,K}^b)} H_{st,K_0} \pi_{s\infty} \pi_{t\infty} \quad \text{and} \quad \Gamma_K^{0b} = [\Gamma_{kl,K}^{0b}],$$

where $I(\mathcal{C}_{k,K}^b)$ denotes a subset of $[K_0]$ such that if $m \in I(\mathcal{C}_{k,K}^b)$, then $\mathcal{C}_{m,K_0} \subset \mathcal{C}_{k,K}^b$.

Assumption 5 *For $K = 2, \cdots, K_0$, $\Gamma_K^{0b} \notin \mathbb{W}_K$, where $\mathbb{W}_K$ is a class of symmetric $K \times K$ matrices which is specified in Appendix D.*

Several remarks are in order. First, the expression of $\mathbb{W}_K$ is complicated and can be found in the proof of Theorem 5 in Appendix D. Second, when $K = 2$,

$$\mathbb{W}_2 = \{W \in \Re^{2\times 2} : W = W^T,\ W_{12}^2 = W_{11}W_{22}\}.$$

In general, we can view $\mathbb{W}_K$ as a set of $K(K+1)/2 \times 1$ vectors. Then, the Lebesgue measure of $\mathbb{W}_K$ is zero, which means Assumption 5 is mild. Third, if the last two columns of Γ_K^{0b} are exactly the same, then $\Gamma_K^{0b} \in \mathbb{W}_K$. Assumption 5 rules out this case when $K \leq K_0$.

Theorem 5 *If Assumptions 1–4 hold, then, for $2 \leq K \leq K_0$, there exists $\tilde{\mathcal{B}}_{K,n}$ such that*

$$\tilde{\varpi}_{K,n}^{-1}\left\{n^{-1}\rho_n^{1/2}[L_n(\hat{Z}_K, \hat{Z}_{K-1}) - \tilde{\mathcal{B}}_{K,n}]\right\} \rightsquigarrow N(0,1)$$

where the asymptotic bias $\tilde{\mathcal{B}}_{K,n}$ and variance $\tilde{\varpi}_{K,n}^2$ are defined in (22) *and* (38)*, respectively, in Appendix D. If, in addition, Assumption 5 holds, then there exist two positive constants (c_{K1}, c_{K2}) potentially dependent on K such that*

$$c_{K2}n^2 \geq \tilde{\mathcal{B}}_{K,n} \geq c_{K1}n^2.$$

Theorem 5 shows that in the under-fitting case, the asymptotic bias term that is of order n^2 will dominate the centered pseudo-LR that is of order $n\rho_n^{-1/2}$. However, when we over-fit the model, i.e., $K > K_0$, the asymptotic bias term will be zero. The sudden change in the orders of magnitude of the pseudo-LR $L_n(\hat{Z}_K^b, \hat{Z}_{K-1})$ provides useful information on the true number of communities.

Next, we consider the over-fitting case. Let z_{K_0+1} be a generic $n \times (K_0+1)$ membership matrix,

$$\begin{aligned} n_{kl}(z_{K_0+1}) &= \sum_{i=1}^{n}\sum_{j\neq i} 1\{[z_{K_0+1}]_{ik} = 1, [z_{K_0+1}]_{jl} = 1\} \\ &= \begin{cases} n_k(z_{K_0+1})n_l(z_{K_0+1}) & \text{if} \quad k \neq l \\ n_k(z_{K_0+1})(n_k(z_{K_0+1}) - 1) & \text{if} \quad k = l, \end{cases} \end{aligned} \tag{6}$$

and $n_k(z_{K_0+1}) = \sum_{l=1}^{K_0+1} n_{kl}(z_{K_0+1})$. We emphasize the dependence of n_{kl} and n_k on the membership matrix z_{K_0+1} because when $K > K_0$, neither Z_K nor Z_K^b is uniquely defined. The following assumption restricts the possible realizations $\hat{Z}_{K_0+1}^b$ can take.

Assumption 6 *There exists some sufficiently small constant ε such that*

$$\inf_{1\leq k\leq K_0+1} n_k(\hat{Z}_{K_0+1}^b)/n \geq \varepsilon.$$

Assumption 6 always holds in our simulation. By Theorem 4, $\hat{Z}_{K_0} = Z_{K_0}$ $a.s.$ Suppose we obtain $\hat{Z}_{K_0+1}^b$ by splitting the last community (i.e., the $\mathcal{C}_{K_0,K_0}$) into two groups by binary segmentation. In simulation, we observe that the two new groups $\widehat{\mathcal{C}}_{K_0,K_0+1}^b$ and $\widehat{\mathcal{C}}_{K_0+1,K_0+1}^b$ have close to even sizes. In addition, we can modify the binary segmentation procedure to ensure that Assumption 6 holds automatically. In particular, suppose $n_{K_0}(\hat{Z}_{K_0+1}^b) \leq n\varepsilon$, then let

$$\widehat{\mathcal{C}}_{K_0,K_0+1}^{b,new} = \widehat{\mathcal{C}}_{K_0,K_0+1}^b \cup \breve{\mathcal{C}}_{K_0+1,K_0+1}^b \quad \text{and} \quad \widehat{\mathcal{C}}_{K_0+1,K_0+1}^{b,new} = \widehat{\mathcal{C}}_{K_0,K_0} \backslash \widehat{\mathcal{C}}_{K_0,K_0+1}^{b,new},$$

where $\breve{\mathcal{C}}^b_{K_0+1,K_0+1}$ is half of $\widehat{\mathcal{C}}^b_{K_0+1,K_0+1}$ by random splitting. Then $\widehat{\mathcal{C}}^{b,new}_{K_0,K_0+1}$ and $\widehat{\mathcal{C}}^{b,new}_{K_0+1,K_0+1}$ satisfy Assumption 6. Although we do not know K_0 a priori, we can apply this modification for any $K = 1, \cdots, K_{\max}$. When $K < K_0$, Theorem 4(2) shows that, for some sufficiently small ε,

$$n_k(\hat{Z}^b_{K+1}) = n_k(Z^b_{K+1}) \geq \inf_k n_{k,K_0} \geq n\varepsilon \quad a.s.$$

Therefore, the modification will never take action when $K < K_0$, which implies that all our results still hold under this modification.

Theorem 6 *Suppose that Assumptions 1–6 hold. Then*

$$0 \leq L_n(\hat{Z}^b_{K_0+1}, \hat{Z}_{K_0}) \leq O_p(n\rho_n^{-1}).$$

In addition, if $h_n \to 0$ and $n\rho_n h_n \to \infty$, then

$$P(\hat{K}_1 \geq K_0) \to 1 \quad \textit{and} \quad P(\hat{K}_2 = K_0) \to 1.$$

Several remarks are in order. First, Theorem 6 establishes the upper bound for the pseudo-LR in the over-fitting case. Like **?**, we are unable to obtain its exact limiting distribution because we do not have the oracle property for $\hat{Z}^b_{K_0+1}$. The more profound reason for the lack of oracle property is that we have limited knowledge on the asymptotic behavior of the (K_0+1)-th column of the eigenvector matrix $\widehat{U}_n$. Fortunately, the upper bound is sufficient for the consistent estimation of K_0 with the help of the tuning parameter h_n. Second, we show that $\hat{K}_1$ cannot under-estimate the number of communities in large samples. This result is similar to that in **?** who showed that NCV does not under-estimate the number of communities in large samples. Third, to obtain a consistent estimate of K_0, we can employ the estimator $\hat{K}_2$ which requires to specify the tuning parameter h_n. This parameter plays the same role as the penalty term in **?**'s BIC-type information criterion. As the average degree $\bar{d}$ is of order $n\rho_n \to \infty$, $h_n = c_h \bar{d}^{-1/2}$ satisfies $h_n \to 0$ and $n\rho_n h_n = c_h(n\rho_n)^{1/2} \to \infty$. Similarly, the average degree is not feasible and is replaced by its sample counterpart in practice. This replacement has theoretical guarantee as discussed after Assumption 4. In Section 4, we investigate the sensitivity of the performance of $\hat{K}_2$ with respect to the constant c_h. We suggest setting $c_h = 1$ based on our simulation results. Last, as mentioned in the introduction, our pseudo-LR method has computational advantages over the existing methods. In particular, it is well known that the likelihood-based method of **?** is computationally expensive even when one uses a variational EM algorithm to approximate the true likelihood. The NCV method of **?** and the ECV method of **?** can also be computationally intensive when the number of folds is large.

4. Numerical Examples on Simulated Networks

4.1 Background and methods

In this section, we conduct simulations to evaluate the performance of our proposed method. We call our pseudo-LR estimators $\widehat{K}_1$ and $\widehat{K}_2$ as PLR1 and PLR2, respectively. Moreover, we compare our proposed method with four other approaches, including LRBIC (**?**), NCV (**?**), ECV (**?**) and BHMC (**?**). LRBIC considers a likelihood-based approach for estimating the latent node labels and selecting models. LRBIC is only designed for the standard SBMs. It requires one to choose a tuning parameter to control the order of the BIC-type penalty. NCV applies cross-validation (CV) from the regularized spectral clustering, while ECV uses CV with edge sampling for choosing between SBM and DCSBM and selecting the number of communities simultaneously. NCV requires one to choose two tuning parameters, viz, the number of folds for the CV and the number of repetitions to reduce the randomness of the estimator due to random sample splitting. ECV requires one to choose two tuning parameters, viz, the probability for an edge to be drawn and the number of replications. BHMC is developed by using the network Bethe-Hessian matrix with moment correction. It requires the selection of a scalar parameter to define the Bethe Hessian matrix and another one for fine-tuning. Like our method, BHMC can be generally applied to both SBM and DCSBM. All methods require to set the maximum number of communities ($K_{\max}$) while searching over K's, and we let $K_{\max} = 10$. We use the R package "randnet" to implement these four methods. The matrix completion procedure for ECV is the default one used in the "randnet" package.

4.2 Data generation mechanisms and settings

We consider the following mechanisms to generate the connectivity matrix $\boldsymbol{B} = \{B_{k\ell}\}_{1\le k,\ell\le K_0}$.

Setting 1 (S1). Let $B_{k\ell} = 0.5\rho n^{-1/2}\{1 + I(k = \ell)\}$ for $1 \le k, \ell \le K_0$, and for some $\rho > 0$.

Setting 2 (S2). Let $B_{k\ell} = 0.9\rho n^{-3/5}\{1 + I(k = \ell)\}$ for $1 \le k, \ell \le K_0$, and for some $\rho > 0$.

Setting 3 (S3). We first simulate $\boldsymbol{W} = (W_1, \ldots, W_{M_0})^\top$ from $\text{Unif}(0, 0.3)^{M_0}$, where $\text{Unif}(a, b)^{M_0}$ denotes an M_0-dimensional uniform distribution on $[a, b]$ and $M_0 = (K_0 + 1)K_0/2$. Let the main diagonal of $\boldsymbol{B}$ be the K_0 largest elements in $\boldsymbol{W}$ and the upper triangular part of $\boldsymbol{B}$ contain the rest elements in $\boldsymbol{W}$. Let $B_{k\ell} = B_{\ell k}$ for all $1 \le k, \ell \le K_0$. We use the generated $\boldsymbol{B}$ with the smallest singular value no smaller than 0.1.

All simulation results are based on 200 realizations. S1 and S2 consider different sparsity levels for different values of ρ, and S3 allows all entries in $\boldsymbol{B}$ to be different. The membership vector is generated by sampling each entry independently from $\{1, \ldots, K_0\}$ with probabilities $\{0.4, 0.6\}$, $\{0.3, 0.3, 0.4\}$ and $\{0.25, 0.25, 0.25, 0.25\}$ for $K_0 = 2, 3$ and 4, respectively. We consider both SBMs and DCSBMs. For the DCSBMs, we generate the degree parameters θ_i from $\text{Unif}(0.2, 1)$ and further normalize them to satisfy the condition (1).

4.3 Results

For our method, we let $\tau = \bar{d}$ and $c_\eta = 0.05$. Note that for computing the PLR2 estimator $\widehat{K}_2$, we need a tuning parameter h_n. We set $h_n = c_h \bar{d}^{-1/2}$. We first would like to examine the performance of the PLR2 estimator when c_h takes different values. Consider $c_h = 0.5, 1.0, 1.5, 2.0$. Let $\rho = 3, 4, 5$ for designs S1 and S2. Tables 1 and 2 report the mean of $\widehat{K}_2$ and $\widehat{K}_1$ by the PLR2 and PLR1 methods, respectively, and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from the DCSBMs of designs S1-S3, for $n = 500, 1000$ and $K_0 = 1, 2, 3, 4$. For saving space, Tables 4 and 5 given in the Supplemental Materials report those statistics when data are generated from the SBMs. The results in Tables 4 and 5 for the SBMs have similar patterns as those in Tables 1 and 2 for the DCSBMs. It is worth noting that when $c_h = 0$, the two estimates $\widehat{K}_1$ and $\widehat{K}_2$ are exactly the same. Comparing Tables 1 and 4 to Tables 2 and 5, we see that for smaller values of c_h, the behavior of $\widehat{K}_2$ is more similar to that of $\widehat{K}_1$. Moreover, Table 1 shows that the PLR2 estimator has similar performance at $c_h = 0.5, 1.0, 1.5, 2.0$ for designs S1 and S2, and its performance improves when the value of ρ or the sample size n increases. However, for design S3, PLR2 behaves better at $c_h = 0.5, 1.0$. Overall, PLR2 at $c_h = 0.5, 1.0$ has good performance for all designs, and PLR2 with $c_h = 1.0$ slightly outperforms PLR1 and PLR2 with $c_h = 0.5$.

Table 1: The mean of $\widehat{K}_2$ and the proportion (prop) of correctly estimating K among 200 simulated datasets when data are generated from DCSBMs.

			$K_0 = 1$				$K_0 = 2$				$K_0 = 3$				$K_4 = 4$			
	ρ	c_h	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
			$n = 500$															
S1	3	mean	1.000	1.000	1.000	1.000	2.095	2.000	2.000	2.000	3.070	3.070	3.000	3.000	3.675	3.675	3.615	3.380
		prop	1.000	1.000	1.000	1.000	0.980	1.000	1.000	1.000	0.980	0.980	1.000	1.000	0.380	0.380	0.390	0.370
	4	mean	1.000	1.000	1.000	1.000	2.035	2.000	2.000	2.000	3.025	3.000	3.000	3.000	4.175	4.150	4.100	4.050
		prop	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	0.990	1.000	1.000	1.000	0.915	0.920	0.935	0.940
	5	mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.020	3.000	3.000	3.000	4.045	4.015	4.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	0.985	0.995	1.000	1.000
S2	3	mean	1.320	1.320	1.320	1.000	2.115	2.000	2.000	2.000	3.090	3.085	3.020	2.990	3.725	3.725	3.680	3.235
		prop	0.880	0.880	0.880	1.000	0.985	1.000	1.000	1.000	0.970	0.975	0.990	0.990	0.290	0.290	0.295	0.280
	4	mean	1.030	1.030	1.000	1.000	2.100	2.000	2.000	2.000	3.030	3.020	3.000	3.000	4.225	4.225	4.150	4.060
		prop	0.995	0.995	1.000	1.000	0.990	1.000	1.000	1.000	0.995	0.995	1.000	1.000	0.915	0.915	0.930	0.945
	5	mean	1.045	1.045	1.000	1.000	2.050	2.000	2.000	2.000	3.025	3.000	3.000	3.000	4.080	4.060	4.040	4.005
		prop	0.995	0.995	1.000	1.000	0.995	1.000	1.000	1.000	0.995	1.000	1.000	1.000	0.970	0.975	0.985	0.990
S3		mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.000	3.000	2.010	2.000	4.000	4.000	3.835	3.665
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.001	0.000	1.000	1.000	0.910	0.825
			$n = 1000$															
S1	3	mean	1.000	1.000	1.000	1.000	2.050	2.000	2.000	2.000	3.000	3.000	3.000	3.000	4.060	4.045	4.025	4.020
		prop	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.980	0.985	0.990	0.995
	4	mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.000	3.000	3.000	3.000	4.020	4.000	4.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000
	5	mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.000	3.000	3.000	3.000	4.020	4.000	4.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000
S2	3	mean	1.045	1.045	1.045	1.000	2.050	2.000	2.000	2.000	3.030	3.030	3.000	3.000	4.225	4.205	4.160	4.020
		prop	0.995	0.995	0.995	1.000	0.990	1.000	1.000	1.000	0.995	0.995	1.000	1.000	0.940	0.945	0.955	0.940
	4	mean	1.040	1.040	1.040	1.000	2.000	2.000	2.000	2.000	3.020	3.000	3.000	3.000	4.015	4.000	4.000	4.000
		prop	0.995	0.995	0.995	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	0.995	1.000	1.000	1.000
	5	mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.015	3.000	3.000	3.000	4.000	4.000	4.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000
S3		mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.000	3.000	3.000	2.030	4.000	4.000	4.000	3.210
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.030	1.000	1.000	1.000	0.605

Based on the above results, we choose $c_h = 1.0$ for the PLR2 estimator. Next, we study the clustering accuracy and the estimation errors of the degree parameters given that the number of communities is correctly selected. We use two commonly used criteria for evaluating the clustering accuracy, which are the Normalized Mutual Information (NMI) calculated from the R package 'aricode' and the percentage (per) of nodes whose memberships are correctly identified. They all give a value between 0 and 1, where 1 means a perfect membership estimation. Moreover, we use the square root of the mean square error (RMSE) for evaluating the estimation accuracy of the degree parameters, defined as $\mathrm{RMSE}_\theta = \sqrt{\sum_i (\hat{\theta}_i - \theta_i)^2/n}$. Table 3 presents the average of the NMI, per and RMSE values given that the number of communities is correctly selected by the PLR2 method based on the 200 realizations for $K_0 = 2, 3, 4$. It also reports the proportion (prop) of correctly

Table 2: The mean of $\widehat{K}_1$ and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from DCSBMs.

			$n = 500$				$n = 1000$			
	ρ		$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$	$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$
S1	3	mean	1.000	2.095	3.070	3.675	1.000	2.050	3.000	4.060
		prop	1.000	0.980	0.980	0.380	1.000	0.990	1.000	0.980
	4	mean	1.000	2.090	3.025	4.175	1.000	2.000	3.000	4.020
		prop	1.000	0.980	0.990	0.915	1.000	1.000	1.000	0.995
	5	mean	1.000	2.035	3.030	4.045	1.000	2.000	3.000	4.045
		prop	1.000	0.990	0.995	0.985	1.000	1.000	1.000	0.985
S2	3	mean	1.320	2.115	3.090	3.725	1.045	2.050	3.030	4.225
		prop	0.880	0.985	0.970	0.290	0.995	0.990	0.995	0.940
	4	mean	1.030	2.100	3.045	4.225	1.040	2.000	3.020	4.015
		prop	0.995	0.990	0.990	0.915	0.995	1.000	0.995	0.995
	5	mean	1.000	2.050	3.025	4.080	1.000	2.000	3.015	4.000
		prop	1.000	0.995	0.990	0.970	1.000	1.000	0.995	1.000
S3		mean	1.000	2.000	3.035	4.005	1.000	2.000	3.000	4.000
		prop	1.000	1.000	0.995	0.995	1.000	1.000	1.000	1.000

estimating K_0 by the PLR2 method. The values of NMI and prop get closer to 1 and the RMSE$_\theta$ values become smaller as the value of ρ or the sample size n increases for all cases. Moreover, we see that K_0 can still be correctly estimated when both clustering errors and estimation errors for θ_i's present. For instance, for design S1 with $n = 500$, the proportions (prop) of correctly estimating K_0 are 100%, 100% and 99.5% for $K_0 = 2$ with $\rho = 3$, $K_0 = 3$ with $\rho = 4$ and $K_0 = 4$ with $\rho = 5$, respectively, while the corresponding NMI values are only 0.873, 0. 849 and 0.822, and the RMSE$_\theta$ values are 0.131, 0.119 and 0.109.

For evaluating the performance of the six methods at different sparsity levels, we let $\rho = 0.5, 1, 2, 3, 4, 5, 6$ for designs S1 and S2. Then the average expected degree ranges from 7.0 to 83.9, for instance, at $K_0 = 4$ and $n = 500$ for the DCSBMs of design S1. Figure 1 shows the proportions of correctly estimating K_0 among 200 simulated datasets versus the values of ρ for the six methods: PLR1 (solid lines), PLR2 (dash-dot lines), LRBIC (dashed lines), NCV (dotted lines), ECV (thin dash-dot lines) and BHMC (thin dotted lines), when data are simulated from the DCSBMs of designs S1 and S2 with $K_0 = 2, 3, 4$ and $n = 500$. The results for designs S1 and S2 are shown in the left and right panels, respectively. The results of the SBMs are presented in Figure 4 given in Appendix B. We observe that our proposed methods PLR1 and PLR2 have similar performance with PLR2 moderately better when $K_0 = 2$. Moreover, PLR1 and PLR2 have larger proportions of correctly estimating K_0 than the other four methods at small values of ρ. This indicates that PLR1 and PLR2 outperform other methods for semi-dense designs. The BHMC method performs better than LRBIC, NCV and ECV at $K_0 = 2, 3$, but its performance becomes inferior to that of the other three methods when $K_0 = 4$. It is worth noting that for larger K_0, it correspondingly requires a larger ρ in order to successfully estimate K_0. When ρ is sufficiently large, eventually all methods can successfully estimate K_0. Compared to the other four methods, PLR1 and PLR2 require less

Table 3: The proportion (prop) of correctly estimating K_0, and the statistics for clustering accuracy (average of NMI and per) and for estimation errors (average of RMSE$_\theta$) given that K_0 is correctly selected by the PLR2 method based on the 200 realizations.

		$K_0 = 2$				$K_0 = 3$				$K_0 = 4$			
	ρ	prop	NMI	per	RMSE$_\theta$	prop	NMI	per	RMSE$_\theta$	prop	NMI	per	RMSE$_\theta$
		$n = 500$											
S1	3	1.000	0.873	0.982	0.131	0.980	0.698	0.918	0.141	0.380	0.488	0.793	0.145
	4	1.000	0.951	0.994	0.110	1.000	0.849	0.966	0.119	0.920	0.685	0.894	0.124
	5	1.000	0.979	0.999	0.096	1.000	0.929	0.986	0.105	0.995	0.822	0.948	0.109
S2	3	1.000	0.861	0.980	0.134	0.975	0.677	0.910	0.144	0.290	0.477	0.787	0.148
	4	1.000	0.945	0.993	0.113	0.995	0.833	0.961	0.122	0.915	0.658	0.881	0.127
	5	1.000	0.976	0.997	0.098	1.000	0.920	0.983	0.107	0.975	0.802	0.941	0.111
S3		1.000	1.000	1.000	0.127	1.000	0.989	0.999	0.150	0.995	0.999	0.999	0.142
		$n = 1000$											
S1	3	1.000	0.950	0.994	0.113	1.000	0.834	0.962	0.121	0.985	0.669	0.890	0.125
	4	1.000	0.986	0.999	0.096	1.000	0.930	0.987	0.103	1.000	0.836	0.955	0.107
	5	1.000	0.996	0.999	0.084	1.000	0.974	0.995	0.091	1.000	0.922	0.981	0.094
S2	3	1.000	0.929	0.991	0.119	1.000	0.786	0.948	0.128	0.945	0.598	0.857	0.132
	4	1.000	0.975	0.997	0.102	1.000	0.902	0.980	0.109	1.000	0.783	0.936	0.113
	5	1.000	0.992	0.999	0.090	1.000	0.958	0.992	0.096	1.000	0.887	0.971	0.100
S3		1.000	1.000	1.000	0.090	1.000	0.999	0.999	0.107	1.000	1.000	1.000	0.094

constraints on the sparsity level ρ in order to correctly estimate K_0. For example, for the DCSBMs of design S1 with $K_0 = 4$, the proportions of correctly estimating K_0 are 0.38 for PLR1 and PLR2, whereas the proportions are close to zero for other methods at $\rho = 3$. For the DCSBMs of design S1 with $K_0 = 2$, the proportions are 0.71 and 0.89 for PLR1 and PLR2, respectively, and they are less than 0.1 for other methods at $\rho = 0.5$.

For further demonstration, Tables 6–8 given in Appendix B report the mean of the estimated number of communities and the proportion (prop) of correctly estimating K_0 for designs S1 and S3 with $n = 500$ obtained from the six methods. Since the results of S2 are similar to those of S1 as shown in Figure 1, we choose to only report those summary statistics for S1. We observe the same pattern as shown in Figure 1 for S1, while the six methods have comparable performance for S3 in which all entries of $\boldsymbol{B}$ are different and the sparsity level is a constant with respect to the sample size.

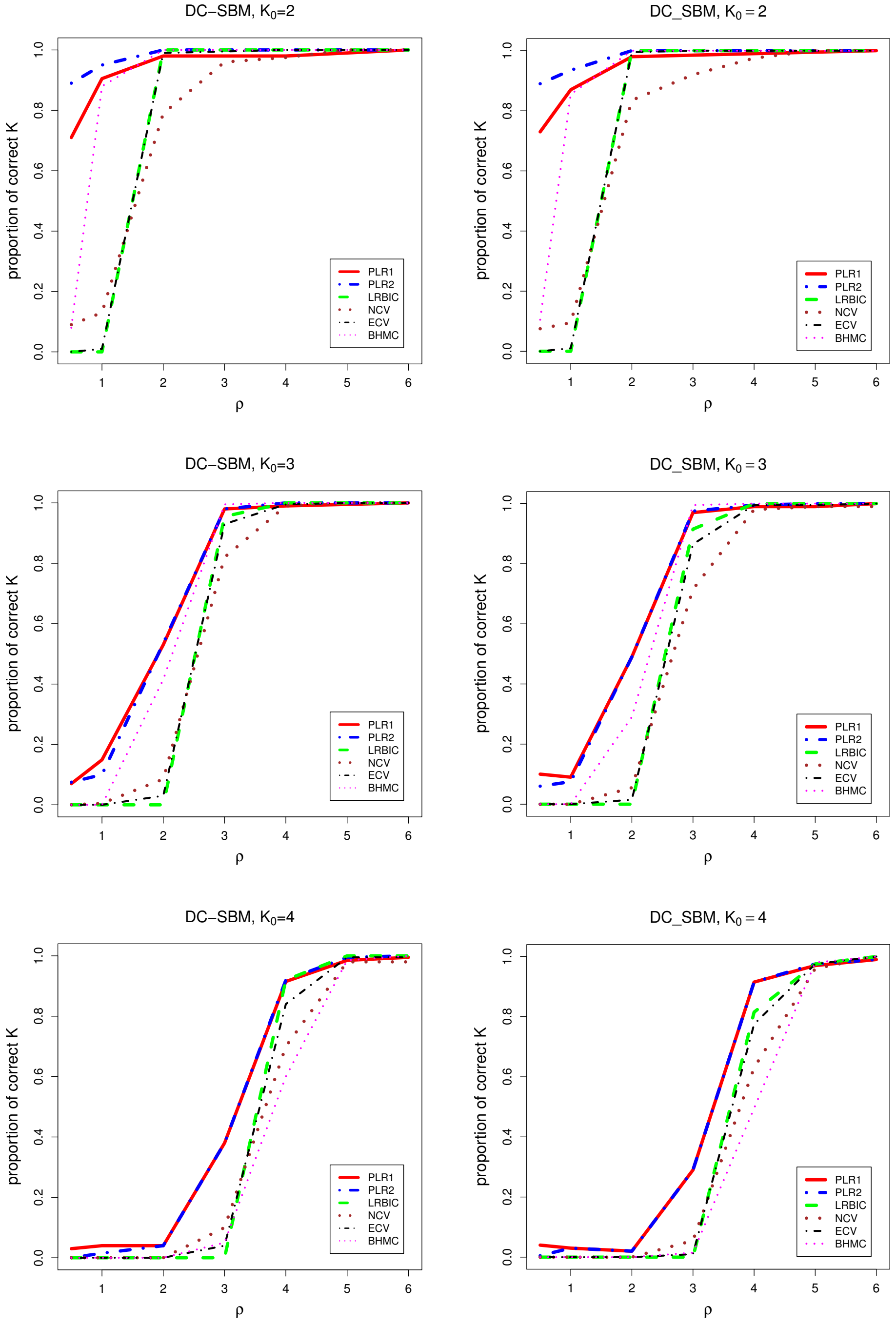
5. Real Data Examples

In this section, we evaluate the performance of our method on several real-world networks.

5.1 Jazz musicians network

We apply the methods to analyze the collaboration network of Jazz musicians. The data are obtained from *The Red Hot Jazz Archive* digital database (www.redhotjazz.com). In our analysis, we include 198 bands that performed between 1912 and 1940. We study the community structure of the band network in which there are 198 nodes representing bands and 2742 unweighted edges indicating at

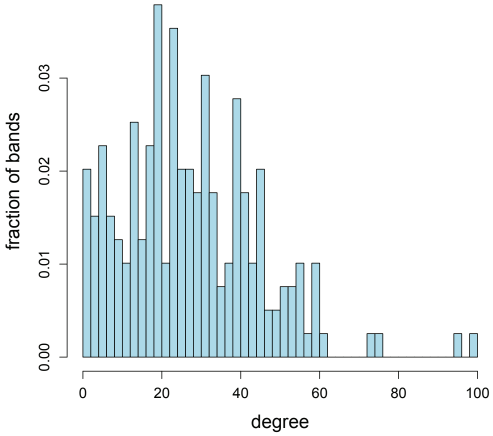
Figure 1: The proportions of correctly estimating K_0 versus the values of ρ for the six methods, when data are simulated from the DCSBMs of designs S1 and S2 with $K_0 = 2, 3, 4$ and $n = 500$. Left panel: design S1; right panel: design S2



(a) design S1

(b) design S2

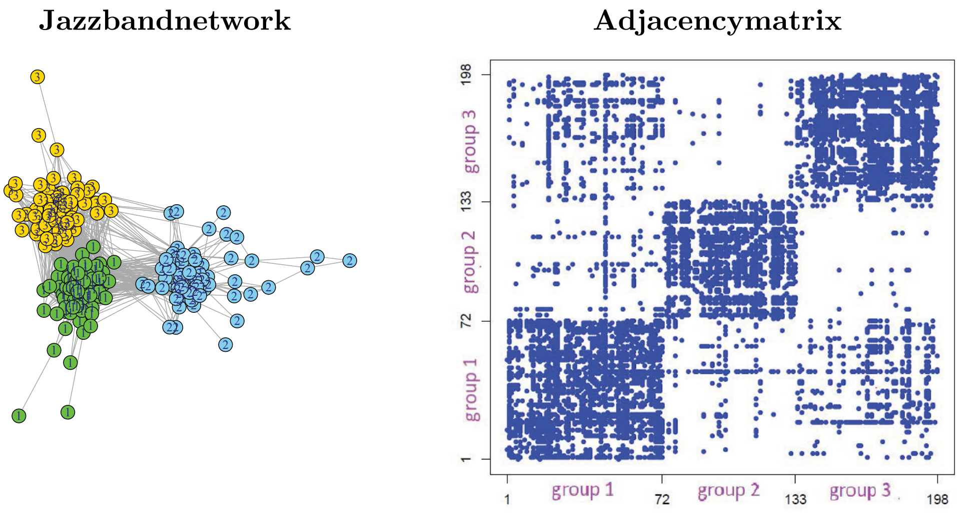
Figure 2: Degree distribution of jazz band network.



least one common musician between two bands. Figure 2 shows the degree distribution for the jazz band network. The minimal, average and maximum degrees of this network are 1.0, 27.7 and 100.0, respectively. Moreover, the distribution of degrees spreads over the range from 1 to 62 with four degree values outside this range. This indicates that the node degrees are highly varying for this network.

Let $K_{\max} = 10$ for all methods. We apply our proposed PLR1 and PLR2 methods to estimate the number of communities and obtain that $\widehat{K}_1 = 3$ and $\widehat{K}_2 = 3$, so that three communities are identified by both methods. For further illustration, the left panel of Figure 3 depicts the band network with 198 nodes divided into three communities. The results confirm the community structure mentioned in **?** that the band network is divided into two large communities based on geographical locations where the bands recorded, and the largest community also splits into two communities due to a racial segregation. Moreover, we obtain the estimated edge probabilities within communities which are $\widehat{B}_{kk} = 0.349, 0.297, 0.358$ for $k = 1, 2, 3$, respectively, and edge probabilities between communities which are $\widehat{B}_{12} = 0.029$, $\widehat{B}_{13} = 0.087$ and $\widehat{B}_{23} = 0.007$. Next, we reorganize the observed adjacency matrix according to the memberships of the nodes, i.e., the nodes in the same estimated community are put together in the adjacency matrix. We use blue dots to represent the edges between nodes. The right panel of Figure 3 shows the reorganized adjacency matrix. We can see that the nodes in the diagonal block matrices, i.e., within each community, are densely connected; while the nodes in the off-diagonal block matrices, i.e., between communities, are sparsely connected. This corroborates the results that the estimated edge probabilities within communities are much larger than those values between communities. Moreover, we observe that the nodes between communities 1 and 3 are more densely connected than the nodes between communities 1 and 3 and communities 2 and 3. This is also consistent with the different estimated edge probabilities between communities that we have obtained. Lastly, we obtain the estimated number of communities as 8, 3, 6 and 7, respectively, by the LRBIC, NCV, ECV and BHMC methods.

Figure 3: left panel depicts the jazz band network with three communities; right panel shows the adjacency matrix reorganized according to the node's memberships.



5.2 Political books network and Facebook friendship network

We apply our methods to a network of US political books (available at www.orgnet.com), and to a large social network which contains friendship data of Facebook users (available at www.snap.stanford.edu). The detailed descriptions of the data applications as well as the numerical results are given in Appendix C.

6. Conclusion

We propose a new pseudo conditional likelihood ratio method for selecting the number of communities in DCSBMs. The method can be naturally applied to SBMs. For estimating the model, we consider the spectral clustering together with a binary segmentation algorithm. This estimation approach enables us to establish the limiting distribution of the pseudo likelihood ratio when the model is under-fitted, and derive the upper bound for it when the model is over-fitted. Based on these properties, we show the consistency of our estimator for the true number of communities. Our method is computationally fast as the estimation is based on spectral clustering, and it also has appealing theoretical properties for the semi-dense and degree-corrected designs. Moreover, our numerical results show that the proposed method has good finite sample performance in various simulation designs and real data applications, and it outperforms several other popular methods in semi-dense networks.

Acknowledgments

The authors are grateful to the Action Editor, Edoardo M. Airoldi, and one anonymous reviewer for their constructive comments that have helped us improve the article substantially. The research of Ma is supported in part by the U.S. NSF grants DMS-17-12558 and DMS-20-14221, and a UCR Academic Senate CoR Grant. Su acknowledges the funding support provided by the Lee Kong Chian Fund for Excellence. Zhang acknowledges the funding support provided by the Singapore Ministry of Education Tier 2 grant under grant no. MOE2018-T2-2-169 and the Lee Kong Chian fellowship.

Appendix A. More details on Algorithms 1 and 2

A.1 Estimators $\hat{P}_{ij}(\hat{Z}_K)$ and $\hat{P}_{ij}(\hat{Z}_K^b)$

By **?**, for a given number of communities K and a generic estimator $\hat{Z}_K$ of the community memberships with corresponding estimated communities $\{\widehat{\mathcal{C}}_{k,K}\}_{k=1}^K$, the maximum likelihood estimators (MLEs) for θ_i and $B_{kl}(\hat{Z}_K)$ in DCSBM are $\hat{\theta}_i = \frac{\hat{d}_i \hat{n}_{k,K}}{\sum_{i' \in \widehat{\mathcal{C}}_{k,K}} \hat{d}_{i'}}$ for $i \in \widehat{\mathcal{C}}_{k,K}$ and $\hat{B}_{kl}(\hat{Z}_K) = \frac{\hat{O}_{kl,K}}{\hat{n}_{kl,K}}$ for $k, l = 1, \cdots, K$, respectively, where $\hat{n}_{k,K} = \sum_{i=1}^n 1\{[\hat{Z}_K]_{ik} = 1\}$,

$$\hat{O}_{kl,K} = \sum_{i=1}^n \sum_{j \neq i} 1\{[\hat{Z}_K]_{ik} = 1, [\hat{Z}_K]_{jl} = 1\} A_{ij}; \tag{7}$$

$$\begin{aligned} \hat{n}_{kl,K} =& \sum_{i=1}^n \sum_{j \neq i} 1\{[\hat{Z}_K]_{ik} = 1, [\hat{Z}_K]_{jl} = 1\} \\ =& \begin{cases} \hat{n}_{k,K}\hat{n}_{l,K} & \text{if} \quad k \neq l \\ \hat{n}_{k,K}(\hat{n}_{k,K} - 1) & \text{if} \quad k = l. \end{cases} \end{aligned} \tag{8}$$

Therefore, for $i \in \widehat{\mathcal{C}}_{k,K}$ and $j \in \widehat{\mathcal{C}}_{l,K}$, when $k \neq l$,

$$\begin{aligned} \hat{P}_{ij}(\hat{Z}_K) =& \hat{\theta}_i \hat{\theta}_j \hat{B}_{kl}(\hat{Z}_K) = \frac{\hat{O}_{kl,K} \hat{d}_i \hat{d}_j}{(\sum_{i' \in \widehat{\mathcal{C}}_{k,K}} \hat{d}_{i'})(\sum_{j' \in \widehat{\mathcal{C}}_{l,K}} \hat{d}_{j'})} \\ =& \frac{\hat{O}_{kl,K} \hat{d}_i \hat{d}_j}{(\sum_{l'=1}^K \hat{O}_{kl',K})(\sum_{l'=1}^K \hat{O}_{ll',K})}; \end{aligned}$$

when $k = l$ and $i, j \in \widehat{\mathcal{C}}_{k,K}$,

$$\hat{P}_{ij}(\hat{Z}_K) = \frac{\hat{O}_{kk,K} \hat{d}_i \hat{d}_j}{\sum_{i',j' \in \widehat{\mathcal{C}}_{k,K}, i' \neq j'} \hat{d}_{i'} \hat{d}_{j'}}.$$

We can compute $\hat{P}_{ij}(\hat{Z}_K^b)$ in the same manner by replacing $\hat{Z}_K$ in the above procedure by $\hat{Z}_K^b$.

A.2 More details on the k-means algorithm

In Algorithm 2, we propose to estimate $\hat{Z}_K$ and $\hat{Z}_{K+1}^b$ by the k-means algorithm. Let $\{\beta_i\}_{i \in \mathcal{C}}$ be a sequence of $d_\beta \times 1$ vectors. The k-means algorithm with K centroids divides $\{\beta_i\}_{i \in \mathcal{C}}$ into K

clusters via solving the following minimization problem:

$$(\alpha_1^*, \cdots, \alpha_K^*) = \arg\min_{\alpha_1,\cdots,\alpha_K} \sum_{i\in\mathcal{C}} \min_{1\leq k\leq K} ||\beta_i - \alpha_k||^2, \tag{9}$$

where the i-th node is classified into cluster k if $k = \arg\min_{1\leq l\leq K} ||\beta_i - \alpha_l^*||$ and if there exists a tie, i.e., $\arg\min_{1\leq l\leq K} ||\beta_i - \alpha_l^*||$ is not a singleton, then we denote k as the smallest minimizer. Then, $\hat{Z}_K$ is obtained by solving (9) with $\beta_i = \hat{\nu}_{iK}$, $i = 1, \cdots, n$ with K centroids. For $\hat{Z}_{K+1}^b$, the binary segmentation step is implemented via solving (9) with 2 centroids and $\beta_i = \hat{\nu}_{iK+1}$, $i \in \widehat{\mathcal{C}}_{k,K}$, for $k = 1, \cdots, K$.

In Section 3.2, we define (Z_K, Z_K^b) by applying Algorithm 2 on ν_{iK}. In view of Theorem 1(2), ν_{iK} takes L_K distinct values $(\bar{\nu}_{1K}, \cdots, \bar{\nu}_{L_K K})$. Let

$$\pi_{l,K} = \#\{i : \nu_{iK} = \bar{\nu}_{lK}\}/n \geq \inf_{1\leq k\leq K_0} \pi_{kn}$$

and g_{iK} be the membership for node i obtained this way, i.e., $g_{iK} = \arg\min_{1\leq k\leq K} ||\nu_{iK} - \alpha_k^*||$ where

$$\begin{aligned}\{\alpha_k^*\}_{k=1}^K =& \arg\min_{\alpha_1,\cdots,\alpha_K} n^{-1} \sum_{i=1}^n \min_{1\leq k\leq K} ||\nu_{iK} - \alpha_k||^2 \\ =& \arg\min_{\alpha_1,\cdots,\alpha_K} \sum_{l=1}^{L_K} \pi_{l,K} \min_{1\leq k\leq K} ||\bar{\nu}_{lK} - \alpha_k||^2. \end{aligned} \tag{10}$$

Then $[Z_K]_{ik} = 1$ if $g_{iK} = k$, $[Z_K]_{ik} = 0$ otherwise, and $\mathcal{C}_{k,K} = \{i : g_{iK} = k\}$. We define Z_{K+1}^b for $K = 1, \cdots, K_0 - 1$ as follows.

1. Given $\{\mathcal{C}_{k,K}\}_{k=1}^K$, let $\widetilde{\mathcal{C}}_{k,K}^l = \mathcal{C}_{k,K} \cap G_{l,K+1}$, for $l = 1, \cdots, L_K$, [5] where $G_{l,K+1}$ is defined in Theorem 1(2). We divide each $\mathcal{C}_{k,K}$ into two subgroups by applying the k-means algorithm to $\{\nu_{iK+1}\}_{i\in\mathcal{C}_{k,K}}$ with two centroids. Denote the two subgroups as $\mathcal{C}_{k,K}(1)$ and $\mathcal{C}_{k,K}(2)$. Note that, by the proof of Theorem 1(2), for $i \in \tilde{\mathcal{C}}_{k,K}^l$, ν_{iK+1} take the same value.

2. For each $k = 1, \cdots, K$, compute

$$Q_K(k) = \frac{\Phi(\mathcal{C}_{k,K}) - \Phi(\mathcal{C}_{k,K}(1)) - \Phi(\mathcal{C}_{k,K}(2))}{\#\mathcal{C}_{k,K}}, \tag{11}$$

where for an arbitrary index set $\mathcal{C}$, $\Phi(\mathcal{C}) = \sum_{i\in\mathcal{C}} ||\nu_{iK+1} - \frac{\sum_{i\in\mathcal{C}} \nu_{iK+1}}{\#\mathcal{C}}||^2$.

3. Choose $k^* = \arg\max_{1\leq k\leq K} Q_K(k)$. Denote

$$\{\mathcal{C}_{k,K+1}^b\}_{k=1}^{K+1} = \{\{\mathcal{C}_{k,K}\}_{k<k^*}, \mathcal{C}_{k^*,K}(1), \{\mathcal{C}_{k,K}\}_{k>k^*}, \mathcal{C}_{k^*,K}(2)\}$$

as the new groups in Z_{K+1}^b.

[5] As can be shown, $\widetilde{\mathcal{C}}_{k,K}^l = G_{l,K+1}$ or $\emptyset$.

Appendix B. Additional simulation results

Tables 4 and 5 given below report the mean of $\widehat{K}_2$ and $\widehat{K}_1$ by the PLR2 and PLR1 methods, respectively, and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from the SBMs of designs S1 and S3 described in Section 4.2, for $n = 500, 1000$ and $K_0 = 1, 2, 3, 4$. Since the results of design S2 have similar patterns as those of S1 when data are generated from the DCSBMs, as shown in Tables 1 and 2, here we choose to only report the results of S1 when data are generated from the SBMs.

Table 4: The mean of $\widehat{K}_2$ and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from SBMs of designs S1 and S3.

			$K_0 = 1$				$K_0 = 2$				$K_0 = 3$				$K_4 = 4$			
	ρ	c_h	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
			$n = 500$															
S1	3	mean	1.035	1.000	1.000	1.000	2.025	2.000	2.000	2.000	3.060	3.060	3.000	3.000	3.465	3.465	3.430	3.355
		prop	0.995	1.000	1.000	1.000	0.995	1.000	1.000	1.000	0.990	0.990	1.000	1.000	0.355	0.355	0.350	0.330
	4	mean	1.000	1.000	1.000	1.000	2.030	2.000	2.000	2.000	3.115	3.015	3.000	3.000	4.085	4.085	4.085	4.005
		prop	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	0.975	0.995	1.000	1.000	0.925	0.925	0.925	0.925
	5	mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.000	3.000	3.000	3.000	4.060	4.060	4.060	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.980	0.980	0.980	1.000
S3		mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.000	3.000	2.035	2.000	4.000	3.995	3.820	3.620
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.035	0.000	1.000	0.995	0.895	0.795
			$n = 1000$															
S1	3	mean	1.000	1.000	1.000	1.000	2.055	2.000	2.000	2.000	3.040	3.005	3.000	3.000	4.080	4.050	4.020	3.990
		prop	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	0.985	0.995	1.000	1.000	0.980	0.990	0.995	0.995
	4	mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.015	3.000	3.000	3.000	4.020	4.000	4.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	0.995	1.000	1.000	1.000
	5	mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.045	3.000	3.000	3.000	4.030	4.020	4.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	0.990	0.995	1.000	1.000
S3		mean	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	3.000	3.000	3.000	2.035	4.000	4.000	4.000	3.320
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.035	1.000	1.000	1.000	0.660

Figure 4 shows the the proportions of correctly estimating K_0 versus the values of ρ for the six methods, PLR1, PLR2, LRBIC, NCV, ECV and BHMC, mentioned in Section 4.1, when the data are simulated from the SBMs of design S1 with $K_0 = 2, 3, 4$ and $n = 500$. We see that our PLR1 and PLR2 outperform the other four methods at small values of ρ. For further comparisons of the six methods, Tables 6-8 report the mean of the estimated number of communities and the proportion (prop) of correctly estimating K_0 for designs S1 and S3 with $n = 500$. For S1, we observe the same pattern as shown in Figures 1 and 4. For S3 in which all entries of $\boldsymbol{B}$ are different, the six methods have comparable performance.

Next, we replace the pseudo likelihood function by the k-means loss function to compare the estimated K communities with the estimated $K+1$ communities obtained from our spectral clustering with binary segmentation method. To this end, we let $Q_n(\hat{Z}^b_{K+1}, \hat{Z}_K)$ be the difference of the k-means loss functions for the estimated K and $K+1$ communities obtained from the first $K+1$ normalized eigenvectors of the regularized graph Laplacian. Then the estimated number of commu-

Table 5: The mean of $\widehat{K}_1$ and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from SBMs of designs S1 and S3.

			$n = 500$				$n = 1000$			
	ρ		$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$	$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$
S1	3	mean	1.035	2.095	3.115	3.465	1.000	2.055	3.040	4.080
		prop	0.995	0.980	0.975	0.355	1.000	0.990	0.985	0.980
	4	mean	1.000	2.045	3.060	4.085	1.000	2.000	3.015	4.020
		prop	1.000	0.990	0.990	0.925	1.000	1.000	0.995	0.995
	5	mean	1.000	2.020	3.015	4.060	1.000	2.000	3.045	4.030
		prop	1.000	0.995	0.995	0.980	1.000	1.000	0.990	0.990
S3		mean	1.000	2.000	3.110	4.000	1.000	2.000	3.000	4.000
		prop	1.000	1.000	0.980	1.000	1.000	1.000	1.000	1.000

Figure 4: The proportions of correctly estimating K_0 versus the values of ρ for the six methods, when data are simulated from the SBMs of design S1 with $K_0 = 2, 3, 4$ and $n = 500$.

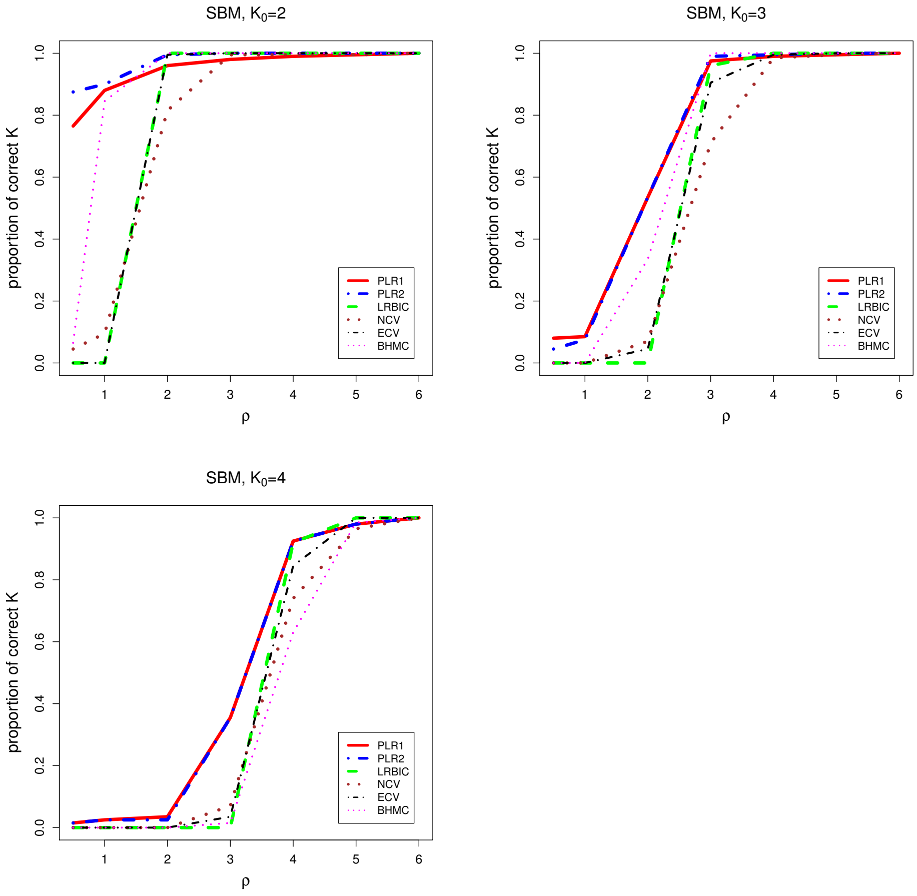


Table 6: The mean of $\widehat{K}$ by the six methods and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets for $K_0 = 2$ and $n = 500$.

		S1							S3
	ρ	0.5	1	2	3	4	5	6	
		SBM							
PLR1	mean	2.865	2.380	2.235	2.095	2.045	2.020	2.000	2.000
	prop	0.765	0.880	0.960	0.980	0.990	0.995	1.000	1.000
PLR2	mean	2.290	2.285	2.025	2.000	2.000	2.000	2.000	2.000
	prop	0.875	0.900	0.995	1.000	1.000	1.000	1.000	1.000
LRBIC	mean	1.000	1.000	2.000	2.000	2.000	2.000	2.000	2.000
	prop	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
NCV	mean	1.055	1.105	2.205	2.005	2.010	2.020	2.000	2.005
	prop	0.045	0.095	0.815	0.995	0.990	0.995	1.000	0.995
ECV	mean	1.000	1.000	2.005	2.000	2.000	2.000	2.000	2.000
	prop	0.000	0.000	0.995	1.000	1.000	1.000	1.000	1.000
BHMC	mean	1.065	1.865	2.000	2.000	2.000	2.000	2.000	2.000
	prop	0.065	0.845	1.000	1.000	1.000	1.000	1.000	1.000
		DCSBM							
PLR1	mean	3.015	2.425	2.120	2.095	2.090	2.035	2.025	2.000
	prop	0.710	0.905	0.980	0.980	0.980	0.990	0.995	1.000
PLR2	mean	2.275	2.205	2.000	2.000	2.000	2.000	2.000	2.000
	prop	0.890	0.950	1.000	1.000	1.000	1.000	1.000	1.000
LRBIC	mean	1.000	1.000	2.000	2.000	2.000	2.000	2.000	2.000
	prop	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
NCV	mean	1.150	1.170	2.040	1.970	1.995	2.000	2.000	2.005
	prop	0.090	0.130	0.790	0.960	0.975	1.000	1.000	0.995
ECV	mean	1.000	1.010	2.000	2.005	2.000	2.000	2.000	2.000
	prop	0.000	0.010	0.990	0.995	1.000	1.000	1.000	1.000
BHMC	mean	1.080	1.880	2.000	2.000	2.000	2.000	2.000	2.000
	prop	0.080	0.880	1.000	1.000	1.000	1.000	1.000	1.000

Table 7: The mean of $\widehat{K}$ by the six methods and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets for $K_0 = 3$ and $n = 500$.

		S1							S3
	ρ	0.5	1	2	3	4	5	6	
		SBM							
PLR1	mean	3.035	2.715	2.975	3.115	3.060	3.015	3.000	3.110
	prop	0.080	0.085	0.535	0.975	0.990	0.995	1.000	0.980
PLR2	mean	2.125	2.595	2.975	3.060	3.015	3.000	3.000	3.000
	prop	0.045	0.075	0.535	0.990	0.995	1.000	1.000	1.000
LRBIC	mean	1.000	1.000	1.005	2.960	3.000	3.000	3.000	3.000
	prop	0.000	0.000	0.000	0.960	1.000	1.000	1.000	1.000
NCV	mean	1.045	1.050	1.495	2.830	3.015	3.015	3.000	3.030
	prop	0.000	0.000	0.070	0.710	0.985	0.995	1.000	0.970
ECV	mean	1.000	1.000	1.400	2.905	3.005	3.000	3.000	3.005
	prop	0.000	0.000	0.045	0.905	0.995	1.000	1.000	0.995
BHMC	mean	1.055	1.160	2.335	3.000	3.000	3.000	3.000	3.000
	prop	0.000	0.000	0.335	1.000	1.000	1.000	1.000	1.000
		DCSBM							
PLR1	mean	2.925	2.930	3.180	3.070	3.025	3.030	3.025	3.035
	prop	0.070	0.149	0.530	0.980	0.990	0.995	0.995	0.995
PLR2	mean	2.125	2.830	3.150	3.070	3.000	3.000	3.000	3.000
	prop	0.075	0.100	0.535	0.980	1.000	1.000	1.000	1.000
LRBIC	mean	1.000	1.000	1.025	2.955	3.000	3.000	3.000	3.000
	prop	0.000	0.000	0.000	0.955	1.000	1.000	1.000	1.000
NCV	mean	1.040	1.065	1.595	2.955	3.000	3.005	3.000	3.010
	prop	0.005	0.000	0.085	0.820	0.990	0.995	1.000	0.990
ECV	mean	1.000	1.000	1.350	2.940	3.005	3.000	3.000	3.000
	prop	0.000	0.000	0.030	0.930	0.995	1.000	1.000	1.000
BHMC	mean	1.055	1.145	2.415	2.995	3.000	3.000	3.000	3.000
	prop	0.000	0.000	0.415	0.995	1.000	1.000	1.000	1.000

Table 8: The mean of $\widehat{K}$ by the six methods and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets for $K_0 = 4$ and $n = 500$.

		S1							S3
	ρ	0.5	1	2	3	4	5	6	
		SBM							
PLR1	mean	2.665	2.850	3.200	3.465	4.085	4.060	4.000	4.000
	prop	0.015	0.025	0.035	0.355	0.925	0.980	1.000	1.000
PLR2	mean	2.300	2.850	2.665	3.465	4.085	4.060	4.000	3.995
	prop	0.015	0.025	0.025	0.355	0.925	0.980	1.000	0.995
LRBIC	mean	1.000	1.000	1.000	1.005	3.840	4.000	4.000	4.000
	prop	0.000	0.000	0.000	0.000	0.920	1.000	1.000	1.000
NCV	mean	1.015	1.020	1.004	1.500	4.030	4.005	4.000	4.060
	prop	0.000	0.000	0.000	0.070	0.740	0.965	1.000	0.940
ECV	mean	1.000	1.000	1.000	1.370	3.905	4.000	4.000	4.000
	prop	0.000	0.000	0.000	0.035	0.845	1.000	1.000	1.000
BHMC	mean	1.035	1.020	1.200	2.330	3.610	3.985	4.000	4.000
	prop	0.000	0.000	0.000	0.015	0.630	0.985	1.000	1.000
		DCSBM							
PLR1	mean	2.750	2.780	2.765	3.675	4.175	4.045	4.010	4.005
	prop	0.030	0.040	0.040	0.380	0.915	0.985	0.995	0.995
PLR2	mean	2.105	2.655	2.745	3.675	4.150	4.015	4.000	4.005
	prop	0.000	0.015	0.040	0.380	0.920	0.995	1.000	0.995
LRBIC	mean	1.000	1.000	1.000	1.005	3.845	4.000	4.000	4.000
	prop	0.000	0.000	0.000	0.000	0.920	1.000	1.000	1.000
NCV	mean	1.050	1.003	1.045	1.805	4.005	4.015	4.020	4.060
	prop	0.000	0.000	0.000	0.100	0.700	0.980	0.980	0.940
ECV	mean	1.000	1.000	1.000	1.435	3.895	4.000	4.005	4.005
	prop	0.000	0.000	0.000	0.040	0.840	1.000	0.995	0.995
BHMC	mean	1.075	1.015	1.285	2.360	3.575	3.985	4.000	4.000
	prop	0.000	0.000	0.000	0.050	0.600	0.985	1.000	1.000

nities minimizes $\frac{Q_n(\hat{Z}^b_{K+1},\hat{Z}_K)/(K+1)}{Q_n(\hat{Z}^b_K,\hat{Z}_{K-1})/K}$, and we call this estimator "KML". Note that $Q_n(\hat{Z}^b_{K+1}, \hat{Z}_K)$ involves the eigenvectors with dimension $n \times (K+1)$. Thus we need to normalize it via dividing it by $K+1$. In addition, we apply the gap statistic proposed in **?** for estimating the number of communities by using the R package "cluster". The gap statistic was proposed for clustering p-dimensional independent vectors into K groups for $K = 1, \cdots, K_{\max}$, where p is fixed and do not change with K. We let $p = K_{\max}$ in our setting, so that we apply this method to the first $K_{\max}$ normalized eigenvectors of the regularized graph Laplacian. Moreover, **?** proposed a semi-definite programming method (SPUR) for determining the number of communities in SBMs. We compare our proposed estimator PLR1 with these three estimators, KML, GAP and SPUR. Since the proposed estimator PLR2 performs slightly better than PLR1, we only compare PLR1 with other three estimators.

Table 9 reports the mean of the estimated number of communities by the four methods, PLR1, KML, GAP and SPUR, and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from the SBMs and designs S1 and S3 given in Section 4.2 with $n = 500$. In Table 10, we report those statistics for the three methods, PLR1, KML, and GAP, when the data are generated from the DCSBMs given in Section 4.2, as the SPUR method was proposed only for the SBMs. Tables 9 and 10 show that our proposed PLR1 has the best performance for all cases. Specifically, the gap statistic method applies the k-means to p-dimensional vectors, where p is fixed and is not allowed to change with K. Hence, it is not directly applicable to network data clustering. As a result, it performs worse than other methods. The KML method performs better than the GAP and SPUR for most cases of design S1, but it is inferior to the proposed PLR1 method, especially for large K's. This is due to the fact that for determining the number of communities, the KML method only uses the information from the eigenvectors, whereas the proposed PLR1 method uses the likelihood which involves all information from the parameter estimates. Moreover, the proposed PLR methods are built on the spectral clustering with binary segmentation algorithm for estimation, and thus they are computationally fast. They have the advantage over the semi-definite programming method, SPUR, in terms of computational speed. Computational efficiency needs to be taken into account for model selection in large network data.

Lastly, for the DCSBMs, we generate the degree parameters θ_i from the Pareto distribution with the scale parameter 1 and the shape parameter 5, and further normalize them to satisfy the condition (1). Tables 11 and 12 report the mean of $\widehat{K}_1$ and $\widehat{K}_2$ with $c_h = 1.0$, respectively, and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets. We see that both PLR1 and PLR2 perform well, and the results in Tables 11 and 12 are comparable to those for $\widehat{K}_1$ and $\widehat{K}_2$ with $c_h = 1.0$ shown in Tables 1 and 2 when θ_i are generated from the uniform distribution.

Appendix C. Additional real data applications

Appendix C.1: Political books network

We investigate the community structure of a network of US political books (available at www.orgnet.com) by different methods. In this network, there are 105 nodes representing books about US politics published around the 2004 presidential election and sold by the online bookseller Amazon.com, and there are 441 edges representing frequent co-purchasing of books by the same buyers. Figure 7 shows the degree distribution for the political books network with the average degree being 8.4. We see that the degree has a right skewed distribution with most values ranging from 2 to

Table 9: The mean of $\widehat{K}$ by the four methods, PLR1, KML, GAP and SPUR, and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from SBMs of S1 and S3 with $n = 500$.

			$K_0 = 2$				$K_0 = 3$				$K_0 = 4$			
	ρ		PLR1	KML	GAP	SPUR	PLR1	KML	GAP	SPUR	PLR1	KML	GAP	SPUR
S1	3	mean	2.095	2.110	7.715	1.815	3.115	2.955	8.615	2.540	3.465	3.155	9.290	3.005
		prop	0.980	0.975	0.115	0.815	0.975	0.895	0.060	0.540	0.355	0.140	0.000	0.115
	4	mean	2.045	2.085	6.265	1.860	3.060	2.965	6.830	2.655	4.085	3.655	8.115	3.515
		prop	0.990	0.980	0.265	0.860	0.990	0.975	0.350	0.655	0.925	0.725	0.115	0.545
	5	mean	2.020	2.040	5.080	1.880	3.015	3.020	5.265	2.755	4.060	3.840	6.320	3.735
		prop	0.995	0.990	0.400	0.880	0.995	0.990	0.610	0.785	0.980	0.900	0.535	0.785
S3		mean	2.000	2.320	9.470	2.000	3.110	3.200	9.265	2.935	4.000	4.000	9.335	3.905
		prop	1.000	0.915	0.000	1.000	0.980	0.970	0.000	0.945	1.000	1.000	0.010	0.925

Table 10: The mean of $\widehat{K}$ by the three methods, PLR1, KML and GAP, and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are generated from DCSBMs of S1 and S3 with $n = 500$.

			$K_0 = 2$			$K_0 = 3$			$K_0 = 4$		
	ρ		PLR1	KML	GAP	PLR1	KML	GAP	PLR1	KML	GAP
S1	3	mean	2.095	2.110	8.210	3.070	2.895	8.855	3.675	3.115	9.300
		prop	0.980	0.975	0.055	0.980	0.875	0.045	0.380	0.135	0.000
	4	mean	2.090	2.095	6.730	3.025	2.955	7.015	4.175	3.525	8.585
		prop	0.980	0.980	0.315	0.990	0.970	0.175	0.915	0.725	0.095
	5	mean	2.035	2.040	5.455	3.030	3.050	6.410	4.045	3.840	6.990
		prop	0.990	0.990	0.490	0.995	0.985	0.420	0.985	0.900	0.410
S2		mean	2.000	2.585	9.375	3.035	3.055	9.440	4.005	4.010	9.455
		prop	1.000	0.850	0.000	0.995	0.990	0.000	0.995	0.990	0.010

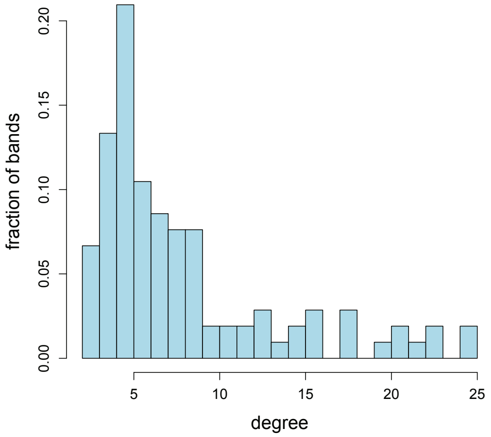
Table 11: The mean of $\widehat{K}_1$ and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are simulated from DCSBMs of S1 and S3 with the degree parameters θ_i generated from the Pareto distribution.

			$n = 500$				$n = 1000$			
	ρ		$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$	$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$
S1	3	mean	1.085	2.095	3.135	3.510	1.000	2.090	3.035	4.045
		prop	0.965	0.985	0.950	0.360	1.000	0.985	0.990	0.990
	4	mean	1.010	2.080	3.040	4.140	1.000	2.050	3.000	4.040
		prop	0.995	0.985	0.990	0.910	1.000	0.990	1.000	0.990
	5	mean	1.000	2.000	3.000	4.045	1.000	2.000	3.000	4.035
		prop	1.000	1.000	1.000	0.985	1.000	1.000	1.000	0.990
S3		mean	1.000	2.000	3.000	4.000	1.000	2.000	3.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 12: The mean of $\widehat{K}_2$ and the proportion (prop) of correctly estimating K_0 among 200 simulated datasets when data are simulated from DCSBMs with the degree parameters θ_i generated from the Pareto distribution.

			$n = 500$				$n = 1000$			
	ρ		$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$	$K_0 = 1$	$K_0 = 2$	$K_0 = 3$	$K_4 = 4$
S1	3	mean	1.085	2.000	3.080	3.510	1.000	2.000	3.015	4.045
		prop	0.965	1.000	0.965	0.360	1.000	1.000	0.995	0.990
	4	mean	1.010	2.000	3.000	4.140	1.000	2.000	3.000	4.040
		prop	0.995	1.000	1.000	0.910	1.000	1.000	1.000	0.990
	5	mean	1.000	2.000	3.000	4.020	1.000	2.000	3.000	4.000
		prop	1.000	1.000	1.000	0.990	1.000	1.000	1.000	1.000
S3		mean	1.000	2.000	3.000	4.000	1.000	2.000	3.000	4.000
		prop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 5: The degree distribution of the political books network.

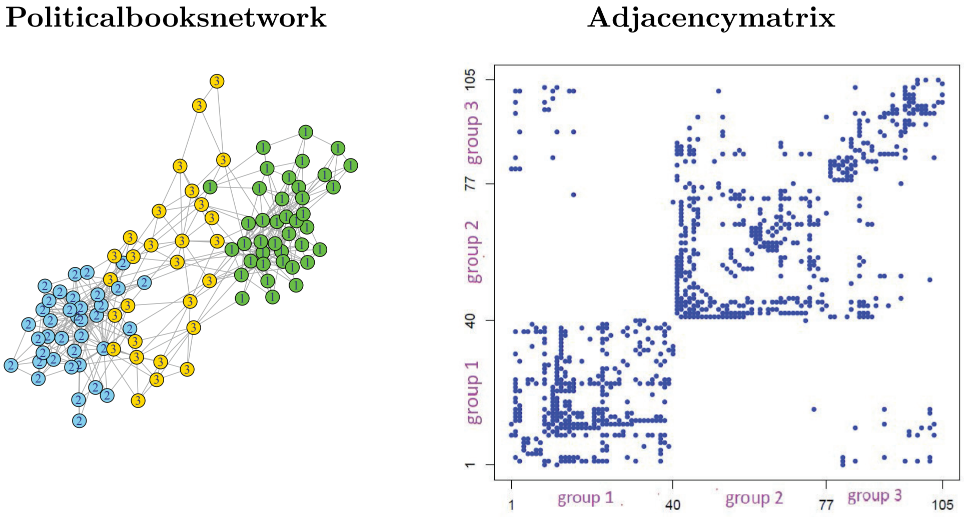


9. Let $K_{\max} = 10$. We identify $\widehat{K}_1 = \widehat{K}_2 = 3$ communities by both PLR1 and PLR2. This result is consistent with the ground-truth community structure that these books are actually divided into three categories "liberal", "neutral" and "conservative" according to their political views (**?**). For further demonstration, we plot the political books network with three communities in the left panel of Figure 6. Groups 1, 2 and 3 represent the estimated communities of liberal, conservative and neutral books. We also obtain the estimated edge probabilities within communities which are $\widehat{B}_{kk} = 0.219, 0.224, 0.164$ for $k = 1, 2, 3$, and the edge probabilities between communities which are $\widehat{B}_{12} = 0.001$, $\widehat{B}_{13} = 0.019$ and $\widehat{B}_{23} = 0.035$. We see that groups 1 and 2 from two different political affiliations are very weakly connected. Moreover, the right panel of Figure 6 shows the adjacency matrix reorganized according to the memberships of the nodes. We use blue dots to represent the edges between nodes. We observe that the nodes are very sparsely connected between communities 1 and 2. The plots in Figure 6 are consistent with the results of the estimated edge probabilities. Lastly, we apply the LRBIC, NCV, ECV and BHMC methods, and obtain the estimated number of communities as 3, 6, 8 and 4, respectively, by these four methods.

Appendix C.2: Facebook friendship network

We apply our methods to a large social network which contains friendship data of Facebook users (available at www.snap.stanford.edu). A node represents a user and an edge represents a friendship between two users. The data have 4039 nodes and 88218 edges. We use the nodes with the degree between 10 and 400. As a result, there are 2901 nodes and 80259 edges in our analysis. The left panel of Figure 8 shows the degree distribution for the Facebook friendship network with the average degree being 55.33. The degree distribution is again right skewed. Let $K_{\max} = 20$. By using the proposed PLR1 and PLR2 methods, we identify $\widehat{K}_1 = \widehat{K}_2 = 11$ communities. The right panel of Figure 8 shows the estimated community structure of the Facebook friendship network with eleven identified communities. We can observe sub-communities of friends who are tightly connected through mutual friendships. Lastly, the LRBIC, NCV, ECV and BHMC methods found 19, 19, 20 and 14 communities, respectively.

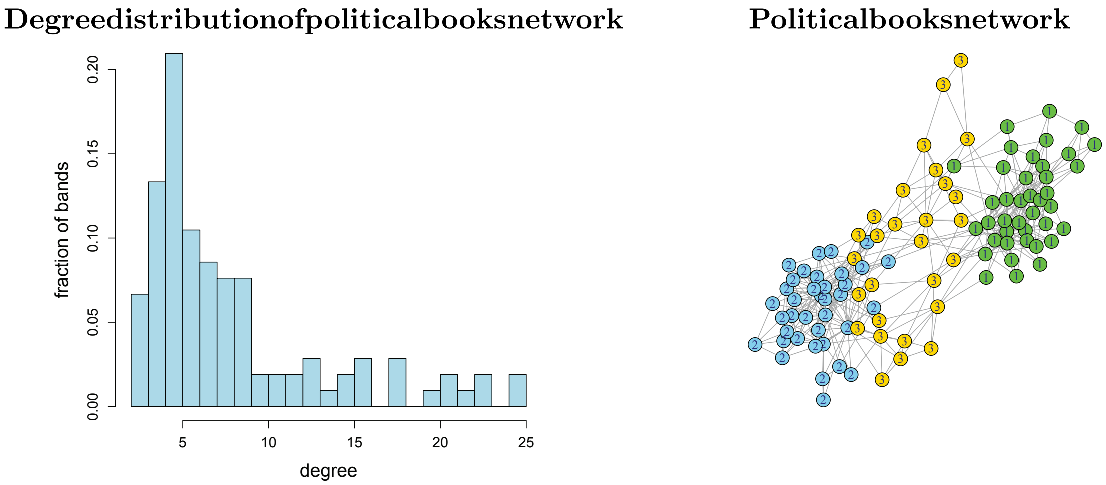
Figure 6: Left panel depicts the political books network with three communities; right panel shows the adjacency matrix reorganized according to the node's memberships.



Politicalbooksnetwork

Adjacencymatrix

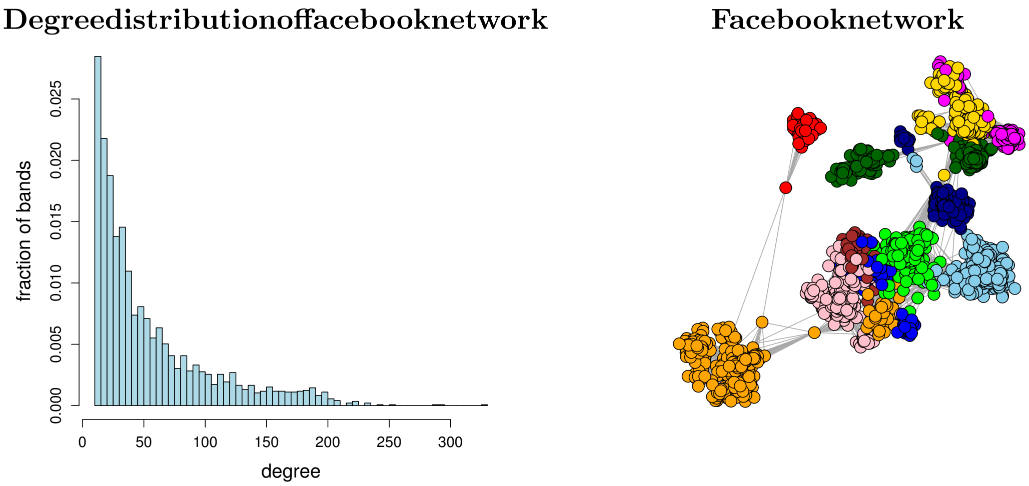
Figure 7: Left panel shows the degree distribution; right panel depicts the political books network with three communities.



Degreedistributionofpoliticalbooksnetwork

Politicalbooksnetwork

Figure 8: Left panel shows the degree distribution; right panel depicts the facebook friendship network with eleven communities.



Degreedistributionoffacebooknetwork

Facebooknetwork

Appendix D. Proofs of results in Section 3

D.1 Proof of Theorem 1

The first result is proved in **?**, Theorem 3.3. For part (2), by Lemma 7(1), if $i \in \mathcal{C}_{k,K_0}$, then

$$u_i^T(K) = (\theta_i^\tau)^{1/2}(n_{k,K_0}^\tau)^{-1/2} S_n^\tau(K).$$

Because $S_n^\tau(K)$ is a $K_0 \times K$ matrix, it is easy to see that $L_K \leq K_0$. By the proof of **?**, Theorem 3.3, S_n^τ is the $K_0 \times K_0$ eigenvector matrix of $(\Pi_n^\tau)^{1/2} H_{0,K_0} (\Pi_n^\tau)^{1/2}$ with the corresponding eigenvalues ordered from the biggest to the smallest in absolute values. By Assumptions 1 and 2, we have

$$(\Pi_n^\tau)^{1/2} H_{0,K_0} (\Pi_n^\tau)^{1/2} \to \Pi_\infty'^{1/2} H_{0,K_0}^* \Pi_\infty'^{1/2} := S_\infty \Sigma_\infty S_\infty.$$

By Davis-Kahan Theorem in **?** and Assumption 2(2), there exists a $K \times K$ orthogonal matrix O_s such that $S_n^\tau(K) O_s \to S_\infty[K]$ where S_∞ is the eigenvector matrix of $\Pi_\infty'^{1/2} H_{0,K_0}^* \Pi_\infty'^{1/2}$ and is of full rank. Therefore, if $i \in \mathcal{C}_{k,K_0}$ and $j \in \mathcal{C}_{l,K_0}$,

$$\begin{aligned} \left\| \frac{u_i^T(K)}{||u_i^T(K)||} - \frac{u_j^T(K)}{||u_j^T(K)||} \right\| &= \left\| \left(\frac{[S_n^\tau]_k(K)}{||[S_n^\tau]_k(K)||} - \frac{[S_n^\tau]_l(K)}{||[S_n^\tau]_l(K)||} \right) O_s \right\| \\ &\to \left\| \frac{[S_\infty]_k(K)}{||[S_\infty]_k(K)||} - \frac{[S_\infty]_l(K)}{||[S_\infty]_l(K)||} \right\|. \end{aligned} \tag{12}$$

Because S_∞ is of full rank, the first K columns of S_∞ should have rank K. This implies the K-dimensional row vectors $\{\frac{[S_\infty]_k(K)}{||[S_\infty]_k(K)||}\}_{k=1}^{K_0}$ take at least K distinct values, which are denoted as $\bar{\nu}_{1,K}, \cdots, \bar{\nu}_{L_K,K}$. Therefore, $L_K \geq K$. Next, we call nodes i and j are equivalent if both $\frac{u_i^T(K)}{||u_i^T(K)||}$

and $\frac{u_j^T(K)}{||u_j^T(K)||}$ converges to one of $(\bar{\nu}_{l,K})$, $l = 1, \cdots, L_K$. Then $G_{l,K}$ can be constructed as the equivalence class of the above equivalence relation. Let

$$I = \left\{(k,l) : \left\| \frac{[S_\infty]_k(K)}{||[S_\infty]_k(K)||} - \frac{[S_\infty]_l(K)}{||[S_\infty]_l(K)||} \right\| \neq 0, k = 1, \cdots, K_0, l = 1, \cdots, K_0 \right\}.$$

In view of the fact that the cardinality of I is finite, we have

$$c^* = \min_{(k,l)\in I} \left\| \frac{[S_\infty]_k(K)}{||[S_\infty]_k(K)||} - \frac{[S_\infty]_l(K)}{||[S_\infty]_l(K)||} = \min_{\ell \neq \ell'} ||\bar{\nu}_{\ell,K} - \bar{\nu}_{\ell',K}|| \right\| > 0.$$

Then, by (12), if nodes $i \notin G_{l,K}$,

$$\liminf_n \left\| \frac{u_i^T(K)}{||u_i^T(K)||} - \bar{\nu}_{l,K} \right\| \geq c^* > 0.$$

This implies that $\{G_{l,K}\}_{l=1}^{L_K}$ constructed as the equivalence class satisfy the two requirements in Theorem 1(2) with $c = c^*$.

D.2 Proof of Theorem 4

First, we prove Theorem 4(1). Let $\hat{g}_{iK}$ be the membership estimated by the k-means algorithm with K centroids, i.e.,

$$\hat{g}_{iK} = \arg\min_{1 \leq k \leq K} ||\hat{\nu}_{iK} - \hat{\alpha}_k|| \quad \text{and} \quad \{\hat{\alpha}_k\}_{k=1}^K = \arg\min_{\alpha_1, \cdots, \alpha_K} \frac{1}{n} \sum_{i=1}^n \min_{1 \leq k \leq K} ||\hat{\nu}_{iK} - \alpha_k||^2.$$

Because the L_2-norm is invariant under rotation,

$$\hat{g}_{iK} = \arg\min_{1 \leq k \leq K} ||\hat{\nu}_{iK}\hat{O}_{Kn}O_s - \hat{\alpha}_k|| \quad \text{and} \quad \{\hat{\alpha}_k\}_{k=1}^K = \arg\min_{\alpha_1, \cdots, \alpha_K} \frac{1}{n} \sum_{i=1}^n \min_{1 \leq k \leq K} ||\hat{\nu}_{iK}\hat{O}_{Kn}O_s - \alpha_k||^2. \tag{13}$$

where $\hat{O}_{Kn}$ is a $K \times K$ orthonormal matrix such that $\hat{O}_{Kn} = \bar{U}\bar{V}^T$, $\bar{U}\bar{\Sigma}\bar{V}^T$ is the singular value decomposition of $\widehat{U}_n(K)^T U_n(K)$, U_n is the population analogue of $\widehat{U}_n$: $\mathcal{L}_\tau = U_n \Sigma_n U_n^T$, and O_s is another $K \times K$ orthonormal matrix defined in the proof of Theorem 1(2). Here, $\Sigma_n = \text{diag}(\sigma_{1n}, \ldots, \sigma_{K_0 n}, 0, ..., 0)$ is a $n \times n$ matrix and we suppress the dependence of $\bar{U}, \bar{\Sigma}$, and $\bar{V}$ on K. We aim to show

$$\sup_i 1\{\hat{g}_{iK} \neq g_{iK}\} = 0 \quad a.s. \tag{14}$$

Suppose that

$$\sup_{1 \leq i \leq n} ||\hat{\nu}_{iK}^T \hat{O}_{Kn} O_s - \nu_{iK}^T|| \leq c_1 \quad a.s., \tag{15}$$

for some sufficiently small $c_1 > 0$, which we will prove later. In addition, by (10),

$$\{\alpha_k^*\}_{k=1}^K = \arg\min_{\alpha_1, \cdots, \alpha_K} \sum_{l=1}^{K_0} \pi_{ln} \min_{1 \leq k \leq K} ||\bar{\nu}_{lK} - \alpha_k||^2.$$

Then for any $k = 1, \cdots, K$, we have

$$\alpha_k^* = \sum_{l \leq K_0 : \mathcal{C}_{l,K_0} \subset \mathcal{C}_{k,K}} \psi_{n,k,l} \bar{\nu}_{lK},$$

or in matrix form,

$$(\alpha_1^*, \cdots, \alpha_K^*) = (\bar{\nu}_{1K}, \cdots, \bar{\nu}_{L_K,K}) \Psi_n',$$

where $\psi_{n,k,l} = \pi_{ln} / (\sum_{l \leq K_0 : \mathcal{C}_{l,K_0} \subset \mathcal{C}_{k,K}} \pi_{ln})$ for $k = 1, \cdots, K$ and $l = 1, \cdots, L_K$, and $\Psi_n = [\psi_{n,k,l}]$. Note that $L_K \geq K$. By Assumption 2, $\Psi_n \to \Psi_\infty$, where $[\Psi_\infty]_{k,l} = \pi_{l\infty} / \sum_{l \leq K_0 : \mathcal{C}_{l,K_0} \subset \mathcal{C}_{k,K}} \pi_{l\infty} > 0$. Because Z_K is unique by Assumption 3(1) and $\pi_{l\infty}$ is positive for $l = 1, \cdots, K_0$, we have that each column of Ψ_∞ has one and only one nonzero entry. In addition, there exist at least $L_K \geq K$ distinct vectors in $\{\bar{\nu}_{lK}\}_{l=1}^{K_0}$. Therefore, by relabeling both $\{\alpha_k^*\}_{k=1}^{K}$ and $\{\bar{\nu}_{lK}\}_{l=1}^{K_0}$, we can make

$$\Psi_\infty' = (\Psi_{1,\infty}, \Psi_{2,\infty}),$$

where $\Psi_{1,\infty}$ is a $K \times K$ diagonal matrix with strictly positive diagonal elements. Therefore, Ψ_∞ has rank K. By Theorem 1(3), $(\bar{\nu}_{1K}, \cdots, \bar{\nu}_{L_K,K})$ also has rank K. This implies, the limit of the $K \times K$ matrix $(\alpha_1^*, \cdots, \alpha_K^*)$ is of full rank. Therefore, there exists a constant $\underline{c} > 0$ such that

$$\liminf_n \min_{k \neq k'} |\alpha_k^* - \alpha_{k'}^*| > \underline{c}. \tag{16}$$

Then (14) follows (15) and Lemma 8(3) with $\hat{\beta}_{in} = \hat{\nu}_{iK} \hat{O}_{Kn} O_s$ and $\beta_{in} = \nu_{iK}$.

Now we turn to prove (15). Since $(\Pi_n^\tau)^{1/2} H_{0,K_0} (\Pi_n^\tau)^{1/2} \to (\Pi_\infty')^{1/2} H_{0,K_0}^* (\Pi_\infty')^{1/2}$ and Assumption 2(2), we have $\inf_n |\sigma_{K+1n} - \sigma_{Kn}| \geq C > 0$ for any $K \leq K_0 - 1$. Second, Assumption 4 implies **?**, Assumption 11. Last, let $d_i^\tau = d_i + \tau$. Since $\tau \leq M n \rho_n$ for some $M > 0$ and $d_i \asymp n\rho_n$, we have,

$$d_i^\tau / d_i \asymp 1.$$

Therefore, there exist constants $C > c > 0$ such that

$$C \geq \sup_{k,n} n_k^\tau d_i^\tau / (n d_i) \geq \inf_{k,n} n_k^\tau d_i^\tau / (n d_i) \geq c.$$

This verifies **?**, Assumption 10. Hence, by **?**, Theorem 3.4,

$$\sup_i (n_{g_{iK_0}}^\tau)^{1/2} \theta_i^{-1/2} ||\hat{u}_i(K)^T \hat{O}_{Kn} - u_i^T(K)|| \leq C^* \log^{1/2}(n) (n\rho_n + \tau)^{-1/2} \leq C^* C_1^{-1/2} \quad a.s., \tag{17}$$

where C^* is a constant independent of n and g_{iK_0} denotes the membership index of node $i,$, viz, $g_{iK_0} = k$ if $[Z_{K_0}]_{ik} = 1$.

In addition, Lemma 7(2) shows that, if $i \in \mathcal{C}_{k,K_0}$ for any $k = 1, \cdots, K_0$, then

$$\liminf_n (n_k^\tau)^{1/2} \theta_i^{-1/2} ||u_i(K)|| = \liminf_n ||[S_n]_k(K)|| \geq c.$$

Therefore,

$$\begin{aligned}
&\sup_i ||\hat{\nu}_{iK}^T \hat{O}_{Kn} O_s - \nu_{iK}^T|| \\
\leq &\sup_i \left\| \hat{\nu}_{iK}^T \hat{O}_{Kn} - \frac{u_i^T(K)}{||u_i(K)||} \right\| + \sup_i \left\| \frac{u_i^T(K) O_s}{||u_i(K)||} - \nu_{iK}^T \right\| \\
\leq &\sup_{1 \leq i \leq n} \frac{||\hat{O}_{Kn}^T \hat{u}_i(K) - u_i(K)||}{||\hat{u}_i(K)||} + o(1) \\
\leq &\frac{C^* C_1^{-1/2}}{c - C^* C_1^{-1/2}} + o(1) \leq c_1, \quad a.s.,
\end{aligned} \tag{18}$$

where the second inequality holds because of the definition of ν_{iK} and Theorem 1. By Assumption 4, C_1 is sufficiently large, which implies that c_1' can be sufficiently small. This concludes the proof of (14).

We also note that, by definition, for any $K = 1, \cdots, K_0$ and $k = 1, \cdots, K_0$, there exists $l = 1, \cdots, L_K$ such that $\mathcal{C}_{k,K_0} \subset G_{l,K}$. In addition, by (10), Assumption 3(1), and Lemma 8(1), for any $l = 1, \cdots, L_K$, there exists $k' = 1, \cdots, K$ such that $G_{l,K} \subset \mathcal{C}_{k',K}$. Therefore,

$$\mathcal{C}_{k,K_0} \subset G_{l,K} \subset \mathcal{C}_{k',K} \quad \text{and} \quad Z_{K_0} \succeq Z_K.$$

Second, we prove Theorem 4(2). We know from Theorem 4(1) that $\hat{Z}_{K-1} = Z_{K-1}$ $a.s.$, i.e., $\widehat{\mathcal{C}}_{k,K-1} = \mathcal{C}_{k,K-1}$ for $k = 1, \cdots, K-1$. We aim to show that $\hat{Z}_K^b = Z_K^b$ $a.s.$ for $K = 2, \cdots, K_0$. Recall $\widetilde{\mathcal{C}}_{k,K-1}^l = \mathcal{C}_{k,K-1} \cap G_{l,K}$. We divide $[K-1]$ into two subsets $\mathcal{K}_1$ and $\mathcal{K}_2$ such that $k \in \mathcal{K}_1$ if there exists at least two indexes l_1 and l_2 such that both $\widetilde{\mathcal{C}}_{k,K-1}^{l_1}$ and $\widetilde{\mathcal{C}}_{k,K-1}^{l_2}$ are nonempty sets and $\mathcal{K}_2 = [K-1] \backslash \mathcal{K}_1$. Note that $L_K \geq K > K-1$. Therefore, by the pigeonhole principle, $\mathcal{K}_1$ is nonempty. We divide the proof into three steps. For a generic $k \in \mathcal{K}_1$, denote $\widehat{\mathcal{C}}_{k,K-1}(1)$ and $\widehat{\mathcal{C}}_{k,K-1}(2)$ as two subsets of $\mathcal{C}_{k,K-1}$ which are obtained by applying k-means algorithm on $\{\hat{\nu}_{in}(K)\}_{i \in \mathcal{C}_{k,K-1}}$ with two centroids. Similarly, let $\mathcal{C}_{k,K-1}(1)$ and $\mathcal{C}_{k,K-1}(2)$ as two subsets of $\mathcal{C}_{k,K-1}$ which are obtained by applying k-means algorithm on $\{\nu_{iK}\}_{i \in \mathcal{C}_{k,K-1}}$ with two centroids. In the first step, we aim to show $\hat{k} = k^* \in \mathcal{K}_1$ $a.s.$, where $\hat{k}$ is defined in Algorithm 2 in Section 2.2. In the second step, we aim to show that $\widehat{\mathcal{C}}_{k^*,K-1}(1) = \mathcal{C}_{k^*,K-1}(1)$ and $\widehat{\mathcal{C}}_{k^*,K-1}(2) = \mathcal{C}_{k^*,K-1}(2)$ $a.s.$ These two results imply that

$$\mathcal{C}_{k^*,K-1}(1) = \widehat{\mathcal{C}}_{\hat{k},K-1}(1) \quad \text{and} \quad \mathcal{C}_{k^*,K-1}(2) = \widehat{\mathcal{C}}_{\hat{k},K-1}(2),$$

which completes the proof of $\hat{Z}_K^b = Z_K^b$ for $k = 1, \cdots, K_0$. Last, in the third step, we show that $Z_{K_0} \succeq Z_{K+1}^b$.

Step 1. We show that $\hat{k} = k^* \in \mathcal{K}_1$ $a.s.$ For a generic $k \in \mathcal{K}_1$, because the L_2-norm is invariant under rotation, we can regard the procedure as applying k-means algorithm to $\hat{\beta}_{in} = O_s^T \hat{O}_{Kn}^T \hat{\nu}_{iK}$

for $i \in \mathcal{C}_{k,K-1}$. Further denote $\beta_{in} = \nu_{iK}$. Then, $\beta_{in} = \beta_{jn}$ if $i, j \in \widetilde{\mathcal{C}}^l_{k,K-1}$ for some l, and

$$\begin{aligned}&\sup_{i \in \mathcal{C}_{k,K-1}} ||\hat{\beta}_{in} - \beta_{in}|| \\ \leq & \sup_{i \in \mathcal{C}_{k,K-1}} \left\| \hat{\nu}^T_{iK} \hat{O}_{Kn} O_s - \frac{u_i^T(K)}{||u_i(K)||} \right\| + \sup_{i \in \ \mathcal{C}_{k,K-1}} \left\| \frac{u_i^T(K) O_s}{||u_i(K)||} - \nu^T_{iK} \right\| \\ \leq & \frac{C^* C_1^{-1/2}}{c - C^* C_1^{-1/2}} + o(1) \leq c_1 \quad a.s.,\end{aligned}$$

where the first inequality holds by the triangle inequality, the second inequality holds because of Theorem 1(2) and the fact that the constant c_1 is sufficiently small. In addition, by the definition of $\{G_{l,K}\}_{l=1}^{L_K}$ in Theorem 1(2), there exists some positive constant c such that, for $l \neq l'$, $\widetilde{\mathcal{C}}^l_{k,K} \neq \emptyset$, and $\widetilde{\mathcal{C}}^{l'}_{k,K} \neq \emptyset$,

$$\inf_{i \in \widetilde{\mathcal{C}}^l_{k,K}, j \in \widetilde{\mathcal{C}}^{l'}_{k,K}} ||\beta_{in} - \beta_{jn}|| \geq c > 0.$$

Recall the definitions of $Q_K(\cdot)$ and $\hat{Q}_K(\cdot)$ in (11) and (5), respectively. Then, by Lemma 8(2), we have, for any $k \in \mathcal{K}_1$, $|Q_{K-1}(k) - \hat{Q}_{K-1}(k)| \leq C' c_1 \ a.s.$ for some constant $C' > 0$. For $k \in \mathcal{K}_2$, $Q_{K-1}(k) = o(1)$ and $|\hat{Q}_{K-1}(k)| \leq C^{''} c_1$. Therefore, $|Q_{K-1}(k) - \hat{Q}_{K-1}(k)| \leq C c_1 \ a.s.$ for $k = 1, \cdots, K-1$. Recall that

$$k^* = \arg\max_{1 \leq k \leq K-1} Q_{K-1}(k)$$

We claim $\hat{k} = k^* \ a.s.$ Suppose not. Then by Assumption 3(2),

$$0 \leq \hat{Q}_{K-1}(\hat{k}) - \hat{Q}_{K-1}(k^*) = Q_{K-1}(\hat{k}) - Q_{K-1}(k^*) + 2C' c_1 \leq 2Cc_1 - c.$$

As c_1 is sufficiently small, we reach a contradiction.

Step 2. We show that $\widehat{\mathcal{C}}_{k^*,K-1}(1) = \mathcal{C}_{k^*,K-1}(1)$ and $\widehat{\mathcal{C}}_{k^*,K-1}(2) = \mathcal{C}_{k^*,K-1}(2) \ a.s.$ Because Z_{K-1} and Z^b_K are unique, Lemma 8(3) implies, up to some relabeling,

$$\mathcal{C}_{k^*,K-1}(1) = \widehat{\mathcal{C}}_{k^*,K-1}(1) \quad \text{and} \quad \mathcal{C}_{k^*,K-1}(2) = \widehat{\mathcal{C}}_{k^*,K-1}(2). \tag{19}$$

Therefore, $\hat{Z}^b_K = Z^b_K$ for $k = 1, \cdots, K_0$.

Step 3. We show that $Z_{K_0} \succeq Z^b_{K+1}$. For any $k = 1, \cdots, K_0$ and any $K = 2, \cdots, K_0$, Theorem 4 (1) shows that there exists $k' \in \{1, \cdots, K-1\}$ such that $\mathcal{C}_{k,K_0} \subset \mathcal{C}_{k',K-1}$. If $k' \neq k^*$, then $\mathcal{C}_{k,K_0} \subset \mathcal{C}_{k',K-1} = \mathcal{C}^b_{k'',K}$ for some $k'' = 1, \cdots, K$. If $k' = k^*$, we know that $\mathcal{C}_{k,K_0} \subset G_{l,K}$ for some $l = 1, \cdots, L_K$. Therefore,

$$\mathcal{C}_{k,K_0} \subset \mathcal{C}_{k^*,K-1} \cap G_{l,K} = \widetilde{\mathcal{C}}^l_{k^*,K-1}.$$

Last, by Lemma 8, we know that

$$\widetilde{\mathcal{C}}^l_{k^*,K-1} \subset \quad \text{either} \quad \mathcal{C}_{k^*,K-1}(1) \quad \text{or} \quad \mathcal{C}_{k^*,K-1}(2).$$

Therefore, there exists $k'' = 1, \cdots, K$ such that

$$\mathcal{C}_{k,K_0} \subset \widetilde{\mathcal{C}}^l_{k^*,K-1} \subset \mathcal{C}^b_{k'',K}.$$

This completes the proof of Theorem 4(2).

For Theorem 4(3), the result holds by the construction of $\hat{Z}^b_{K+1}$ for $K = 1, \cdots, K_0$ and the fact that $\hat{Z}_K = Z_K$ for $K = 1, \cdots, K_0$.

D.3 Proof of Theorem 5

We first state $\mathbb{W}_K$: if $K = 2$,

$$\mathbb{W}_K = \left\{ \begin{array}{c} W \in \Re^{K\times K} : W \text{ is symmetric}, \\ W_{K-1K-1}(W_{K-1\cdot} + W_{K\cdot})^2 = W_{K-1\cdot}^2(W_{K-1K-1} + 2W_{K-1K} + W_{KK}), \\ W_{K-1K}(W_{K-1\cdot} + W_{K\cdot})^2 = W_{K-1\cdot}W_{K\cdot}(W_{K-1K-1} + 2W_{K-1K} + W_{KK}), \\ W_{KK}(W_{K-1\cdot} + W_{K\cdot})^2 = W_{K\cdot}^2(W_{K-1K-1} + 2W_{K-1K} + W_{KK}), \\ \end{array} \right\}$$

and if $K \geq 3$

$$\mathbb{W}_K = \left\{ \begin{array}{c} W \in \Re^{K\times K} : W \text{ is symmetric}, \\ W_{kl}(W_{K-1\cdot} + W_{K\cdot}) = W_{l\cdot}(W_{kK-1} + W_{kK}),\ k = 1, \cdots, K-2,\ l = K-1, K, \\ W_{K-1K-1}(W_{K-1\cdot} + W_{K\cdot})^2 = W_{K-1\cdot}^2(W_{K-1K-1} + 2W_{K-1K} + W_{KK}), \\ W_{K-1K}(W_{K-1\cdot} + W_{K\cdot})^2 = W_{K-1\cdot}W_{K\cdot}(W_{K-1K-1} + 2W_{K-1K} + W_{KK}), \\ W_{KK}(W_{K-1\cdot} + W_{K\cdot})^2 = W_{K\cdot}^2(W_{K-1K-1} + 2W_{K-1K} + W_{KK}), \\ \end{array} \right\}$$

where $W_{k\cdot} = \sum_{l=1}^{K} W_{kl}$ for $W = [W_{kl}] \in \Re^{K\times K}$.

By Theorem 4, we have $\hat{Z}_K^b = Z_K^b$ $a.s.$ for $K \leq K_0$. By Theorem 4(3), without loss of generality, we assume that $\hat{Z}_K^b = Z_K^b$ is obtained by splitting the last group in $\hat{Z}_{K-1} = Z_{K-1}$ into the $(K-1)$-th and K-th groups in $\hat{Z}_K$, i.e.,

$$\#\mathcal{C}_{k,K-1} = \#\mathcal{C}_{k,K}^b, \text{ for } k = 1, \cdots, K-2 \quad \text{and} \quad \#\mathcal{C}_{K-1,K-1} = \#\mathcal{C}_{K-1,K}^b \cup \#\mathcal{C}_{K,K}^b.$$

Define $O_{kl,K}^b$ and $O_{kl,K}$ as (7) with $\hat{Z}_K$ replaced by Z_K^b and Z_K, respectively, and $n_{kl,K}^b$ and $n_{kl,K}$ as (8) with $\hat{Z}_K$ replaced by Z_K^b and Z_K, respectively. Further define

$$\widehat{M}_{kl,K} = \frac{O_{kl,K}}{(\sum_{l'=1}^{K} O_{kl',K})(\sum_{l'=1}^{K} O_{ll',K})} \quad \text{and} \quad \widehat{M}_{kl,K}^b = \frac{O_{kl,K}^b}{(\sum_{l'=1}^{K} O_{kl',K}^b)(\sum_{l'=1}^{K} O_{ll',K}^b)}, \quad k \neq l,$$

$$\widehat{M}_{kk,K} = \frac{O_{kk,K}}{\sum_{i,j\in \mathcal{C}_{k,K}, i\neq j} \hat{d}_i \hat{d}_j}, \quad \text{and} \quad \widehat{M}_{kk,K}^b = \frac{O_{kk,K}^b}{\sum_{i,j\in \mathcal{C}_{k,K}^b, i\neq j} \hat{d}_i \hat{d}_j}.$$

Then, almost surely, for $i \in \widehat{\mathcal{C}}_{k,K}$ and $i \in \widehat{\mathcal{C}}_{l,K}$

$$\hat{P}_{ij}(\hat{Z}_K) = \widehat{M}_{kl,K}\hat{d}_i\hat{d}_j,$$

and for $i \in \widehat{\mathcal{C}}_{k,K}^b$ and $i \in \widehat{\mathcal{C}}_{l,K}^b$

$$\hat{P}_{ij}(\hat{Z}_K^b) = \widehat{M}_{kl,K}^b\hat{d}_i\hat{d}_j.$$

Then, for any $k, l \leq K-2$, if $i \in \mathcal{C}_{k,K}^b = \mathcal{C}_{k,K-1}$ and $j \in \mathcal{C}_{l,K}^b = \mathcal{C}_{l,K-1}$, we have

$$O_{kl,K}^b = O_{kl,K-1}, \sum_{i'\in\mathcal{C}_{k,K}^b} \hat{d}_{i'} = \sum_{i'\in\mathcal{C}_{k,K-1}} \hat{d}_{i'}, \quad \text{and thus,} \quad \hat{P}_{ij}(\hat{Z}_K^b) = \hat{P}_{ij}(\hat{Z}_{K-1}).$$

By (2),

$$\begin{aligned}
&L_n(\hat{Z}_K^b, \hat{Z}_{K-1})\\
=&2\sum_{k=1}^{K-2}\bigg\{\sum_{l=K-1}^{K} 0.5n_{kl,K}^b\bigg(\frac{\widehat{M}_{kl,K}^b}{\widehat{M}_{kK-1,K-1}}-1\bigg)^2\bigg\}\\
&+\bigg\{0.5\bigg[n_{K-1K-1,K}^b\bigg(\frac{\widehat{M}_{K-1K-1,K}^b}{\widehat{M}_{K-1K-1,K-1}}-1\bigg)^2\\
&+2n_{K-1K,K}^b\bigg(\frac{\widehat{M}_{K-1K,K}^b}{\widehat{M}_{K-1K-1,K-1}}-1\bigg)^2+n_{KK,K}^b\bigg(\frac{\widehat{M}_{KK,K}^b}{\widehat{M}_{K-1K-1,K-1}}-1\bigg)^2\bigg]\bigg\}\\
=:&2\sum_{k=1}^{K-2}\hat{I}_{kn}+\widehat{II}_n.
\end{aligned}$$

For $i \in \mathcal{C}_{k,K}^b$ and $j \in \mathcal{C}_{l,K}^b$, $k,l = 1,\cdots,K$, the population counterparts of $\hat{P}_{ij}(\hat{Z}_K)$ and $\hat{P}_{ij}(\hat{Z}_K^b)$ are

$$P_{ij}(Z_K) = \frac{E[O_{kl,K}]d_i d_j}{\sum_{i'\in\mathcal{C}_{k,K},j'\in\mathcal{C}_{l,K},i'\neq j'} d_{i'}d_{j'}} := M_{kl,K}^b d_i d_j \tag{20}$$

and

$$P_{ij}(Z_K^b) = \frac{E[O_{kl,K}^b]d_i d_j}{\sum_{i'\in\ \mathcal{C}_{k,K}^b,j'\in\mathcal{C}_{l,K}^b,i'\neq j'} d_{i'}d_{j'}} := M_{kl,K}^b d_i d_j, \tag{21}$$

respectively. Let

$$\tilde{\mathcal{B}}_{K,n} = 2\sum_{k=1}^{K-2} I_{kn} + II_n, \tag{22}$$

where

$$\begin{aligned}
I_{kn} &= \sum_{l=K-1}^{K} 0.5n_{kl,K}^b\bigg(\frac{M_{kl,K}^b}{M_{kK-1,K-1}}-1\bigg)^2 \text{ and} &(23)\\
II_n &= 0.5n_{K-1K-1,K}^b\bigg(\frac{M_{K-1K-1,K}^b}{M_{K-1K-1,K-1}}-1\bigg)^2\\
&\quad+n_{K-1K,K}^b\bigg(\frac{M_{K-1K,K}^b}{M_{K-1K-1,K-1}}-1\bigg)^2+0.5n_{KK,K}^b\bigg(\frac{M_{KK,K}^b}{M_{K-1K-1,K-1}}-1\bigg)^2. &(24)
\end{aligned}$$

Note that $O_{kl,K}^b$ is independent across $1 \le k,l \le K$. Let

$$V_{kl,K}^b = \frac{\sum_{s\in I(\mathcal{C}_{k,K}^b),t\in I(\mathcal{C}_{l,K}^b)}[n_\theta^{(1)}(s,t)H_{st,K_0} - n_\theta^{(2)}(s,t)H_{st,K_0}B_{st}(Z_{K_0})]}{n^2},$$

where $n_\theta^{(m)}(k) = \sum_{i\in\mathcal{C}_{k,K_0}}\theta_i^m$ for $m = 1,\cdots,4$,

$$n_\theta^{(1)}(s,t) = n_\theta^{(1)}(s)n_\theta^{(1)}(t) - n_\theta^{(2)}(s)1\{s=t\},$$

and

$$n_\theta^{(2)}(s,t) = n_\theta^{(2)}(s)n_\theta^{(2)}(t) - n_\theta^{(4)}(s)1\{s=t\}.$$

Then,

$$n^{-1}\rho_n^{-1/2}\{O^b_{kl,K} - E[O^b_{kl,K}]\} - N_K(k,l) = o_p(1), \quad k \neq l, \tag{25}$$

where $N_K(k,l)$ is normally distributed with expectation zero and variance $V^b_{kl,K}$,

$$n^{-1}\rho_n^{-1/2}\{O^b_{kk,K} - E[O^b_{kk,K}]\} - N_K(k,k) = o_p(1), \quad k = K-1, K,$$

where $N_K(k,k)$ is normally distributed with zero expectation and variance $2V^b_{kk,K}$, and

$$\{\{N_K(k,l)\}_{k=1,\cdots,K-2,l=K-1,K}, N_K(K-1,K), N_K(K-1,K-1), N_K(K,K)\}$$

are mutually independent.

Next, we consider the linear expansions for $\hat{I}_{kn} - I_{kn}$ and $\widehat{II}_n - II_n$ separately in Steps 1 and 2 below.

Step 1. We consider the linear expansion of $\hat{I}_{kn} - I_{kn}$.
In this step, we focus on the case in which $k = 1, \cdots, K-2$ and $l = K-1, K$. Note that

$$\begin{aligned}\frac{\widehat{M}^b_{kl,K}}{\widehat{M}_{kK-1,K-1}} =& \frac{O^b_{kl,K}/[\sum_{l'=1}^K O^b_{ll',K}]}{O_{kK-1,K-1}/[\sum_{l'=1}^{K-1} O_{K-1l',K-1}]} \\ =& \frac{O^b_{kl,K}/[\sum_{l'=1}^K O^b_{ll',K}]}{[\sum_{l=K-1}^K O^b_{kl,K}]/[\sum_{l=K-1}^K \sum_{l'=1}^K O^b_{ll',K}]}.\end{aligned}$$

Similarly,

$$\frac{M^b_{kl,K}}{M_{kK-1,K-1}} = \frac{E[O^b_{kl,K}]/\{\sum_{l'=1}^K E[O^b_{ll',K}]\}}{\{\sum_{l=K-1}^K E[O^b_{kl,K}]\}/\{\sum_{l=K-1}^K \sum_{l'=1}^K E[O^b_{ll',K}]\}}. \tag{26}$$

Then, by the delta method and some tedious calculation, we have

$$n\rho_n^{1/2}[\widehat{M}^b_{kl,K} - M^b_{kl,K}] = \frac{N_K(k,l)}{\Gamma^b_{l\cdot,K}} - \frac{\Gamma^b_{kl,K}[\sum_{l'=1}^K N_K(l,l')]}{(\Gamma^b_{l\cdot,K})^2} + o_p(1),$$

where $N_K(K-1,K) = N_K(K,K-1)$,

$$\Gamma^b_{kl,K} = n^{-2}\rho_n^{-1}E[O_{kl}] = \Gamma^{0b}_{kl,K} + o(1), \tag{27}$$

and

$$\Gamma^b_{l\cdot,K} = n^{-2}\rho_n^{-1}\sum_{l'=1}^K E[O^b_{ll',K}] = \Gamma^{0b}_{l\cdot,K} + o(1). \tag{28}$$

Similarly,

$$\begin{aligned}&n\rho_n^{1/2}[\widehat{M}_{kK-1,K-1} - M_{kK-1,K-1}] \\ =&\frac{N_K(k,K-1) + N_K(k,K)}{\Gamma^b_{K-1\cdot,K} + \Gamma^b_{K\cdot,K}} \\ &- \frac{[\Gamma^b_{kK-1,K} + \Gamma^b_{kK,K}][\sum_{l'=1}^K N_K(l',K-1) + N_K(l',K)]}{[\Gamma^b_{K-1\cdot,K} + \Gamma^b_{K\cdot,K}]^2} + o_p(1).\end{aligned}$$

By Taylor expansion, we have

$$
\begin{aligned}
& n\rho_n^{1/2}\bigg(\frac{\widehat{M}^b_{kl,K}}{\widehat{M}_{kK-1,K-1}}-\frac{M^b_{kl,K}}{M_{kK-1,K-1}}\bigg)\\
=&\frac{1}{M_{kK-1,K-1}}\bigg[\frac{N_K(k,l)}{\Gamma^b_{l\cdot,K}}-\frac{\Gamma^b_{kl,K}(\sum_{l'=1}^K N_K(l,l'))}{(\Gamma^b_{l\cdot,K})^2}\bigg]\\
&-\frac{M^b_{kl,K}}{M^2_{kK-1,K-1}}\bigg[\frac{N_K(k,K-1)+N_K(k,K)}{\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}}\\
&-\frac{(\Gamma^b_{kK-1,K}+\Gamma^b_{kK,K})(\sum_{l'=1}^K N_K(l',K-1)+N_K(l',K))}{(\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K})^2}\bigg]+o_p(1).
\end{aligned}
$$

This, in conjunction with the fact that $a^2-b^2=(a-b)^2+2\,(a-b)\,b$, implies that

$$
\begin{aligned}
& n^{-1}\rho_n^{1/2}(\hat{I}_{kn}-I_{kn}) \qquad (29)\\
=&\sum_{l=K-1}^{K}0.5n^{-1}\rho_n^{1/2}n^b_{kl,K}\bigg(\frac{\widehat{M}^b_{kl,K}}{\widehat{M}_{kK-1,K-1}}-\frac{M^b_{kl,K}}{M_{kK-1,K-1}}\bigg)^2\\
&+\sum_{l=K-1}^{K}n^{-1}\rho_n^{1/2}n^b_{kl,K}\bigg(\frac{\hat{M}^b_{kl,K}}{\widehat{M}_{kK-1,K-1}}-\frac{M^b_{kl,K}}{M_{kK-1,K-1}}\bigg)\bigg(\frac{M^b_{kl,K}}{M_{kK-1,K-1}}-1\bigg)\\
=&\sum_{l=K-1}^{K}\pi^b_{k,K}\pi^b_{l,K}\bigg(\frac{M^b_{kl,K}}{M_{kK-1,K-1}}-1\bigg)\\
&\times n\rho_n^{1/2}\bigg(\frac{\widehat{M}^b_{kl,K}}{\widehat{M}_{kK-1,K-1}}-\frac{M^b_{kl,K}}{M_{kK-1,K-1}}\bigg)+o_p(1)\\
=&\sum_{l'=1}^{K-2}\sum_{l=K-1}^{K}\phi_{l',l}(k)N_K(l',l)+\phi_{K-1,K-1}(k)N_K(K-1,K-1)+\phi_{K-1,K}(k)N_K(K-1,K)\\
&+\phi_{K,K}(k)N_K(K,K)+o_p(1),
\end{aligned}
$$

where the second equality follows from the facts that $n^b_{kl,K}=n^b_{k,K}n^b_{l,K}$, $n^b_{k,K}=\sum_{i=1}^n 1\{[Z^b_K]_{ik}=1\}$, and

$$
\frac{n^b_{k,K}}{n}\to\pi^b_{k,K}:=\sum_{m\in I(\mathcal{C}^b_{k,K})}\pi_{m\infty}
$$

with $\pi_{m\infty}$ defined in Assumption 2 and that $n\rho_n^{1/2} \to \infty$ as $n \to \infty$ under Assumption 4. For the last line of the above display,

$$\begin{aligned}
&\phi_{l',l}(k)\\
=&\pi^b_{k,K}\pi^b_{l,K}\left(\frac{M^b_{kl,K}}{M^2_{kK-1,K-1}} - \frac{1}{M_{kK-1,K-1}}\right)\left[\frac{1\{l'=k\}}{\Gamma^b_{l\cdot,K}} - \frac{\Gamma^b_{kl,K}}{(\Gamma^b_{l\cdot,K})^2}\right]\\
&-\sum_{l=K-1}^{K}\pi^b_{k,K}\pi^b_{l,K}\left(\frac{(M^b_{kl,K})^2}{M^3_{kK-1,K-1}} - \frac{M^b_{kl,K}}{M^2_{kK-1,K-1}}\right)\\
&\times\left[\frac{1\{l'=k\}}{\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}} - \frac{\Gamma^b_{kK-1,K}+\Gamma^b_{kK,K}}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2}\right], \quad l'=1,\cdots,K-2, \quad l=K-1,K,
\end{aligned}$$

$$\begin{aligned}
&\phi_{K-1,K-1}(k)\\
=&-\pi^b_{k,K}\pi^b_{K-1,K}\left(\frac{M^b_{kK-1,K}}{M^2_{kK-1,K-1}} - \frac{1}{M_{kK-1,K-1}}\right)\frac{\Gamma^b_{kK-1,K}}{(\Gamma^b_{K-1\cdot,K})^2}\\
&+\sum_{l=K-1}^{K}\pi^b_{k,K}\pi^b_{l,K}\left(\frac{(M^b_{kl,K})^2}{M^3_{kK-1,K-1}} - \frac{M^b_{kl,K}}{M^2_{kK-1,K-1}}\right)\frac{\Gamma^b_{kK-1,K}+\Gamma^b_{kK,K}}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2},
\end{aligned}$$

$$\begin{aligned}
&\phi_{K-1,K}(k)\\
=&-\sum_{l=K-1}^{K}\pi^b_{k,K}\pi^b_{l,K}\left(\frac{M^b_{kl,K}}{M^2_{kK-1,K-1}} - \frac{1}{M_{kK-1,K-1}}\right)\frac{\Gamma^b_{kl,K}}{(\Gamma^b_{l\cdot,K})^2}\\
&+\sum_{l=K-1}^{K}\pi^b_{k,K}\pi^b_{l,K}\left(\frac{(M^b_{kl,K})^2}{M^3_{kK-1,K-1}} - \frac{M^b_{kl,K}}{M^2_{kK-1,K-1}}\right)\frac{2[\Gamma^b_{kK-1,K}+\Gamma^b_{kK,K}]}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2},
\end{aligned}$$

and

$$\begin{aligned}
&\phi_{K,K}(k)\\
=&-\pi^b_{k,K}\pi^b_{K,K}\left(\frac{M^b_{kK,K}}{M^2_{kK-1,K-1}} - \frac{1}{M_{kK-1,K-1}}\right)\frac{\Gamma^b_{kK,K}}{(\Gamma^b_{K\cdot,K})^2}\\
&+\sum_{l=K-1}^{K}\pi^b_{k,K}\pi^b_{l,K}\left(\frac{(M^b_{kl,K})^2}{M^3_{kK-1,K-1}} - \frac{M^b_{kl,K}}{M^2_{kK-1,K-1}}\right)\frac{\Gamma^b_{kK-1,K}+\Gamma^b_{kK,K}}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2}.
\end{aligned}$$

Step 2. We consider the linear expansion of $\widehat{II}_n - II_n$.

Note that

$$\begin{aligned}
&\widehat{M}^b_{K-1K-1,K} - M^b_{K-1K-1,K}\\
=&\frac{O^b_{K-1K-1,K} - E[O^b_{K-1K-1,K}]}{\sum_{i',j'\in\mathcal{C}^b_{K-1,K},i'\neq j'}\hat{d}_{i'}\hat{d}_{j'}}\\
&-\frac{E[O^b_{K-1K-1,K}][\sum_{i',j'\in\mathcal{C}^b_{K-1,K},i'\neq j'}(\hat{d}_{i'}\hat{d}_{j'} - d_{i'}d_{j'})]}{(\sum_{i',j'\in\mathcal{C}^b_{K-1,K},i'\neq j'}\hat{d}_{i'}\hat{d}_{j'})(\sum_{i',j'\in\mathcal{C}^b_{K-1,K},i'\neq j'}d_{i'}d_{j'})}.
\end{aligned}$$

By the proof of **?**, Lemma 3.1, we have, for some positive constant $C > 0$,

$$\sup_i |\hat{d}_i/d_i - 1| \leq C(\log^{1/2}(n)(n\rho_n)^{-1/2}) \leq CC_1^{-1/2} \quad a.s. \tag{30}$$

Therefore,

$$\begin{aligned}
n^{-4}\rho_n^{-2} \sum_{i',j' \in \mathcal{C}^b_{K-1,K}, i' \neq j'} \hat{d}_{i'}\hat{d}_{j'} =& n^{-4}\rho_n^{-2} \left[(\sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}_{i'})^2 - \sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}^2_{i'} \right] \\
=& n^{-4}\rho_n^{-2} \left[\left(\sum_{k=1}^{K} (EO^b_{kK-1,K} + O^b_{kK-1,K} - EO^b_{kK-1,K}) \right)^2 - \sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}^2_{i'} \right] \\
=& [\Gamma^b_{K-1\cdot,K} + O_p((n\rho_n^{1/2})^{-1})]^2 - n^{-4}\rho_n^{-2} \sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}^2_{i'} \\
=& (\Gamma^b_{K-1\cdot,K})^2 + o_p(1),
\end{aligned}$$

where the third equality holds because $O^b_{kK-1,K} - EO^b_{kK-1,K} = O_p(n\rho_n^{1/2})$ and the last equality holds because

$$n^{-4}\rho_n^{-2} \sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}^2_{i'} \leq n^{-4}\rho_n^{-2} \sum_{i' \in \mathcal{C}^b_{K-1,K}} d_i^2(1 + CC_1^{-1/2}) = O_{a.s.}(n^{-1}).$$

Also note that, by (30),

$$\begin{aligned}
& n^{-3}\rho_n^{-3/2} \sum_{i',j' \in \mathcal{C}^b_{K-1,K}, i' \neq j'} (\hat{d}_{i'}\hat{d}_{j'} - d_{i'}d_{j'}) \\
=& n^{-3}\rho_n^{-3/2} \left[(\sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}_{i'})^2 - (\sum_{i' \in \mathcal{C}^b_{K-1,K}} d_{i'})^2 \right] - n^{-3}\rho_n^{-3/2} \left[\sum_{i' \in \mathcal{C}^b_{K-1,K}} (\hat{d}^2_{i'} - d^2_{i'}) \right] \\
=& n^{-3}\rho_n^{-3/2} \left[(\sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}_{i'} - d_{i'})(\sum_{i' \in \mathcal{C}^b_{K-1,K}} d_{i'} + \hat{d}_{i'}) \right] + o_{a.s.}(1) \\
=& n^{-3}\rho_n^{-3/2} \left[(\sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}_{i'} - d_{i'}) 2 (\sum_{i' \in \mathcal{C}^b_{K-1,K}} d_{i'}) \right] + n^{-3}\rho_n^{-3/2} (\sum_{i' \in \mathcal{C}^b_{K-1,K}} \hat{d}_{i'} - d_{i'})^2 \\
& + o_{a.s.}(1) \\
=& 2\Gamma_{K-1\cdot,K} \left(\sum_{l'=1}^{K} N_K(K-1, l') \right) + o_p(1),
\end{aligned}$$

where the second equality holds because

$$
\begin{aligned}
&n^{-3}\rho_n^{-3/2}\left|\sum_{i'\in\mathcal{C}^b_{K-1,K}}(\hat{d}^2_{i'}-d^2_{i'})\right| \\
=&n^{-3}\rho_n^{-3/2}\left|\sum_{i'\in\mathcal{C}^b_{K-1,K}}(\hat{d}_{i'}-d_{i'})\right|\left|\sum_{i'\in\mathcal{C}^b_{K-1,K}}(\hat{d}_{i'}+d_{i'})\right| \\
\le&n^{-3}\rho_n^{-3/2}\left(1+CC_1^{-1/2}\right)\left[\sum_{i'\in\mathcal{C}^b_{K-1,K}}d_{i'}\right]^2 C(\log^{1/2}(n)(n\rho_n)^{-1/2})=o_{a.s.}(1),
\end{aligned}
$$

and the last equality holds because

$$\sum_{i'\in\mathcal{C}^b_{K-1,K}}(\hat{d}_{i'}-d_{i'})=O_p(n\rho_n^{1/2}).$$

Then, by the delta method,

$$
\begin{aligned}
&n^3\rho_n^{3/2}[\widehat{M}^b_{K-1K-1,K}-M^b_{K-1K-1,K}] \\
=&\frac{N_K(K-1,K-1)}{(\Gamma^b_{K-1\cdot,K})^2}-\frac{2\Gamma^b_{K-1K-1,K}[\sum_{l'=1}^K N_K(K-1,l')]}{(\Gamma^b_{K-1\cdot,K})^3}+o_p(1).
\end{aligned}
\tag{31}
$$

Similarly,

$$n^3\rho_n^{3/2}(\widehat{M}^b_{KK,K}-M^b_{KK,K})=\frac{N_K(K,K)}{(\Gamma^b_{K\cdot,K})^2}-\frac{2\Gamma^b_{KK,K}[\sum_{l'=1}^K N_K(K,l')]}{(\Gamma^b_{K\cdot,K})^3}+o_p(1).$$

Furthermore, we have

$$
\begin{aligned}
&\widehat{M}^b_{K-1K,K}-M^b_{K-1K,K} \\
=&\frac{O^b_{K-1K,K}-E[O^b_{K-1K,K}]}{(\sum_{i'\in\mathcal{C}^b_{K-1,K}}\hat{d}_{i'})(\sum_{j'\in\mathcal{C}^b_{K,K}}\hat{d}_{j'})} \\
&-\frac{E[O^b_{K-1K,K}][(\sum_{i'\in\mathcal{C}^b_{K-1,K}}\hat{d}_{i'})(\sum_{j'\in\mathcal{C}^b_{K,K}}\hat{d}_{j'})-(\sum_{i'\in\mathcal{C}^b_{K-1,K}}d_{i'})(\sum_{j'\in\mathcal{C}^b_{K,K}}d_{j'})]}{(\sum_{i'\in\mathcal{C}^b_{K-1,K}}\hat{d}_{i'})(\sum_{j'\in\mathcal{C}^b_{K,K}}\hat{d}_{j'})(\sum_{i'\in\mathcal{C}^b_{K-1,K}}d_{i'})(\sum_{j'\in\mathcal{C}^b_{K,K}}d_{j'})}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&n^3\rho_n^{3/2}[\widehat{M}^b_{K-1K,K}-M^b_{K-1K,K}] \\
=&\frac{N_K(K-1,K)}{\Gamma^b_{K-1\cdot,K}\Gamma^b_{K\cdot,K}} \\
&-\frac{\Gamma^b_{K-1K,K}[\Gamma^b_{K-1\cdot,K}\sum_{l'=1}^K N_K(l',K)+\Gamma^b_{K\cdot,K}\sum_{l'=1}^K N_K(l',K-1)]}{(\Gamma^b_{K-1\cdot,K})^2(\Gamma^b_{K\cdot,K})^2}+o_p(1).
\end{aligned}
\tag{32}
$$

Finally, noting that

$$\begin{aligned}
&\widehat{M}_{K-1K-1,K-1} \\
=&\frac{O_{K-1K-1,K-1}}{\sum_{i',j'\in\mathcal{C}_{K-1,K-1},i'\neq j'}\hat{d}_{i'}\hat{d}_{j'}} \\
=&\frac{O^b_{K-1K-1,K}+2O^b_{K-1K,K}+O^b_{KK,K}}{\sum_{i',j'\in\mathcal{C}^b_{K-1,K},i'\neq j'}\hat{d}_{i'}\hat{d}_{j'}+\sum_{i',j'^b_{K,K},i'\neq j'}\hat{d}_{i'}\hat{d}_{j'}+2\sum_{i'\in\mathcal{C}^b_{K-1,K},j'\in\mathcal{C}^b_{K,K}}\hat{d}_{i'}\hat{d}_{j'}},
\end{aligned}$$

we have

$$\begin{aligned}
&n^3\rho_n^{3/2}(\widehat{M}_{K-1K-1,K-1}-M_{K-1K-1,K-1}) \\
=&\frac{N_K(K-1,K-1)+2N_K(K-1,K)+N_K(K,K)}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2} \\
&-\frac{\Gamma^b_{K-1K-1,K}+2\Gamma^b_{K-1K,K}+\Gamma^b_{KK,K}}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^3} \\
&\times\left\{\sum_{l'=1}^{K}2[N_K(K-1,l')+N_K(K,l')]\right\}+o_p(1).
\end{aligned}\tag{33}$$

For $s,t=K-1,K$, let $\hat{m}^b_{st,K}=n^2\rho_n\widehat{M}^b_{st,K}$ and

$$m^b_{st,K}=n^2\rho_n M^b_{st,K}=\frac{\Gamma^{0b}_{st,K}}{\Gamma^{0b}_{s\cdot,K}\Gamma^{0b}_{t\cdot,K}}[1+o(1)].$$

Define $m_{K-1K-1,K-1}$ and $\hat{m}_{K-1K-1,K-1}$ similarly. By the previous calculations, we have

$$\hat{m}^b_{st,K}=m^b_{st,K}[1+o_{a.s.}(1)].$$

Hence,

$$\begin{aligned}
&n\rho_n^{1/2}\left(\frac{\widehat{M}^b_{K-1K-1,K}}{\widehat{M}_{K-1K-1,K-1}}-\frac{M^b_{K-1K-1,K}}{M_{K-1K-1,K-1}}\right) \\
=&\frac{n^3\rho_n^{3/2}[\widehat{M}^b_{K-1K-1,K}-M^b_{K-1K-1,K}]}{m_{K-1K-1,K-1}} \\
&-\frac{m^b_{K-1K-1,K}n^3\rho_n^{3/2}[\widehat{M}_{K-1K-1,K-1}-M_{K-1K-1,K-1}]}{m^2_{K-1K-1,K-1}}+o_p(1),
\end{aligned}\tag{34}$$

$$\begin{aligned}
&n\rho_n^{1/2}\left(\frac{\widehat{M}^b_{KK,K}}{\widehat{M}_{K-1K-1,K-1}}-\frac{M^b_{KK,K}}{M_{K-1K-1,K-1}}\right) \\
=&\frac{n^3\rho_n^{3/2}[\widehat{M}^b_{KK,K}-M^b_{KK,K}]}{m_{K-1K-1,K-1}} \\
&-\frac{m^b_{KK,K}n^3\rho_n^{3/2}[\widehat{M}_{K-1K-1,K-1}-M_{K-1K-1,K-1}]}{m^2_{K-1K-1,K-1}}+o_p(1),
\end{aligned}\tag{35}$$

and

$$\begin{aligned}
&n\rho_n^{1/2}\bigg(\frac{\widehat{M}^b_{K-1K,K}}{\widehat{M}_{K-1K-1,K-1}}-\frac{M^b_{K-1K,K}}{M_{K-1K-1,K-1}}\bigg)\\
=&\frac{n^3\rho_n^{3/2}[\widehat{M}^b_{K-1K,K}-M^b_{K-1K,K}]}{m_{K-1K-1,K-1}}\\
&-\frac{m^b_{K-1K,K}n^3\rho_n^{3/2}[\widehat{M}_{K-1K-1,K-1}-M_{K-1K-1,K-1}]}{m^2_{K-1K-1,K-1}}+o_p(1). \qquad (36)
\end{aligned}$$

Then, by (31)–(36),

$$\begin{aligned}
&n^{-1}\rho_n^{1/2}(\widehat{II}_n-II_n) \qquad (37)\\
=&n\rho_n^{1/2}\bigg[(\pi^b_{K-1,K})^2\bigg(\frac{\widehat{M}^b_{K-1K-1,K}}{\widehat{M}_{K-1K-1,K-1}}-\frac{M^b_{K-1K-1,K}}{M_{K-1K-1,K-1}}\bigg)\\
&+2\pi^b_{K-1,K}\pi^b_{K,K}\bigg(\frac{\widehat{M}^b_{K-1K,K}}{\widehat{M}_{K-1K-1,K-1}}-\frac{M^b_{K-1K,K}}{M_{K-1K-1,K-1}}\bigg)\\
&+(\pi^b_{K,K})^2\bigg(\frac{\widehat{M}^b_{KK,K}}{\widehat{M}_{K-1K-1,K-1}}-\frac{M^b_{KK,K}}{M_{K-1K-1,K-1}}\bigg)\bigg]+o_p(1)\\
=&n^3\rho_n^{3/2}\bigg[\frac{(\pi^b_{K-1,K})^2[\widehat{M}^b_{K-1K-1,K}-M^b_{K-1K-1,K}]}{m_{K-1K-1,K-1}}\\
&+\frac{2\pi^b_{K-1,K}\pi^b_{K,K}[\widehat{M}^b_{K-1K,K}-M^b_{K-1K,K}]}{m_{K-1K-1,K-1}}\\
&+\frac{(\pi^b_{K,K})^2[\widehat{M}^b_{KK,K}-M^b_{KK,K}]}{m_{K-1K-1,K-1}}\bigg]\\
&+\frac{(\pi^b_{K-1,K})^2m^b_{K-1K-1,K}+2\pi^b_{K-1,K}\pi^b_{K,K}m^b_{K-1K,K}+(\pi^b_{K,K})^2m^b_{KK,K}}{m^2_{K-1K-1,K-1}}\\
&\times n^3\rho_n^{3/2}[\widehat{M}_{K-1K-1,K-1}-M_{K-1K-1,K-1}]+o_p(1)\\
=&\sum_{l'=1}^{K-2}\sum_{l=K-1}^{K}\phi_{l',l}(K-1)N_K(l',l)+\phi_{K-1,K-1}(K-1)N_K(K-1,K-1)\\
&+\phi_{K-1,K}(K-1)N_K(K-1,K)+\phi_{K,K}(K-1)N_K(K,K)+o_p(1),
\end{aligned}$$

where, by denoting

$$\phi=\frac{(\pi^b_{K-1,K})^2m^b_{K-1K-1,K}+2\pi^b_{K-1,K}\pi^b_{K,K}m^b_{K-1K,K}+(\pi^b_{K,K})^2m^b_{KK,K}}{m^2_{K-1K-1,K-1}},$$

we have

$$\begin{aligned}
&\phi_{l',K-1}(K-1)\\
=&-\frac{2(\pi^b_{K-1,K})^2\Gamma^b_{K-1K-1,K}}{(\Gamma^b_{K-1\cdot,K})^3 m_{K-1K-1,K-1}}-\frac{2\pi^b_{K-1,K}\pi^b_{K,K}\Gamma^b_{K-1K,K}}{\Gamma^b_{K\cdot,K}(\Gamma^b_{K-1\cdot,K})^2 m_{K-1K-1,K-1}}\\
&-\frac{2\phi[\Gamma^b_{K-1K-1,K}+2\Gamma^b_{K-1K,K}+\Gamma^b_{KK,K}]}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^3 m^2_{K-1K-1,K-1}},\quad l'=1,\cdots,K-2,
\end{aligned}$$

$$\begin{aligned}
&\phi_{l',K}(K-1)\\
=&-\frac{2(\pi^b_{K,K})^2\Gamma^b_{KK,K}}{(\Gamma^b_{K\cdot,K})^3 m_{K-1K-1,K-1}}-\frac{2\pi^b_{K-1,K}\pi^b_{K,K}\Gamma^b_{K-1K,K}}{(\Gamma^b_{K\cdot,K})^2\Gamma^b_{K-1\cdot,K} m_{K-1K-1,K-1}}\\
&-\frac{2\phi[\Gamma^b_{K-1K-1,K}+2\Gamma^b_{K-1K,K}+\Gamma^b_{KK,K}]}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^3 m^2_{K-1K-1,K-1}},\quad l'=1,\cdots,K-2,
\end{aligned}$$

$$\begin{aligned}
&\phi_{K-1,K-1}(K-1)\\
=&\frac{(\pi^b_{K-1,K})^2}{(\Gamma^b_{K-1\cdot,K})^2 m_{K-1K-1,K-1}}-\frac{2(\pi^b_{K-1,K})^2\Gamma^b_{K-1K-1,K}}{(\Gamma^b_{K-1\cdot,K})^3 m_{K-1K-1,K-1}}-\frac{2\pi^b_{K-1,K}\pi^b_{K,K}\Gamma^b_{K-1K,K}}{\Gamma^b_{K\cdot,K}(\Gamma^b_{K-1\cdot,K})^2 m_{K-1K-1,K-1}}\\
&+\frac{\phi}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2 m^2_{K-1K-1,K-1}}-\frac{2\phi[\Gamma^b_{K-1K-1,K}+2\Gamma^b_{K-1K,K}+\Gamma^b_{KK,K}]}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^3 m^2_{K-1K-1,K-1}},
\end{aligned}$$

$$\begin{aligned}
&\phi_{K,K}(K-1)\\
=&\frac{(\pi^b_{K,K})^2}{(\Gamma^b_{K\cdot,K})^2 m_{K-1K-1,K-1}}-\frac{2(\pi^b_{K,K})^2\Gamma^b_{KK,K}}{(\Gamma^b_{K\cdot,K})^3 m_{K-1K-1,K-1}}-\frac{2\pi^b_{K-1,K}\pi^b_{K,K}\Gamma^b_{K-1K,K}}{(\Gamma^b_{K\cdot,K})^2\Gamma^b_{K-1\cdot,K} m_{K-1K-1,K-1}}\\
&+\frac{\phi}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2 m^2_{K-1K-1,K-1}}-\frac{2\phi[\Gamma^b_{K-1K-1,K}+2\Gamma^b_{K-1K,K}+\Gamma^b_{KK,K}]}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^3 m^2_{K-1K-1,K-1}},
\end{aligned}$$

and

$$\begin{aligned}
&\phi_{K-1,K}(K-1)\\
=&-\frac{2(\pi^b_{K-1,K})^2\Gamma^b_{K-1K-1,K}}{(\Gamma^b_{K-1\cdot,K})^3 m_{K-1K-1,K-1}}-\frac{2(\pi^b_{K,K})^2\Gamma^b_{KK,K}}{(\Gamma^b_{K\cdot,K})^3 m_{K-1K-1,K-1}}+\frac{2\pi^b_{K-1,K}\pi^b_{K,K}}{\Gamma^b_{K-1\cdot,K}\Gamma^b_{K\cdot,K} m_{K-1K-1,K-1}}\\
&-\frac{2\pi^b_{K-1,K}\pi^b_{K,K}\Gamma^b_{K-1K,K}[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]}{(\Gamma^b_{K\cdot,K})^2(\Gamma^b_{K-1\cdot,K})^2 m_{K-1K-1,K-1}}+\frac{2\phi}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^2 m^2_{K-1K-1,K-1}}\\
&-\frac{4\phi[\Gamma^b_{K-1K-1,K}+2\Gamma^b_{K-1K,K}+\Gamma^b_{KK,K}]}{[\Gamma^b_{K-1\cdot,K}+\Gamma^b_{K\cdot,K}]^3 m^2_{K-1K-1,K-1}}.
\end{aligned}$$

Combining (29) and (37), we have

$$
\begin{aligned}
& n^{-1}\rho_n^{1/2}[L_n(\hat{Z}_K, \hat{Z}_{K-1}) - \tilde{\mathcal{B}}_{K,n}] \\
= & \sum_{l'=1}^{K-2} \sum_{l=K-1}^{K} \phi_{l',l} N_K(l', l) + \phi_{K-1,K-1} N_K(K-1, K-1) \\
& + \phi_{K-1,K} N_K(K-1, K) + \phi_{K,K} N_K(K, K) + o_p(1),
\end{aligned}
$$

where

$$
\phi_{l',l} = \sum_{k=1}^{K-2} 2\phi_{l',l}(k) + \phi_{l',l}(K-1), \quad l' = 1, \cdots, l, \quad l = K-1,\ K.
$$

Letting

$$
\begin{aligned}
\tilde{\varpi}_{K,n}^2 = & \sum_{l'=1,\cdots,K-2;\, l=K-1,K;\, l' \le l} \phi_{l',l}^2 V_{l'l,K}^b + \phi_{K-1,K-1}^2 2V_{K-1K-1,K}^b \\
& + \phi_{K,K}^2 2V_{KK,K}^b + \phi_{K-1,K}^2 V_{K-1K,K}^b,
\end{aligned} \tag{38}
$$

we have

$$
\tilde{\varpi}_{K,n}^{-1} \left\{ n^{-1}\rho_n^{1/2}[L_n(\hat{Z}_K, \hat{Z}_{K-1}) - \tilde{\mathcal{B}}_{K,n}] \right\} \rightsquigarrow N(0,1).
$$

Step 3. We now prove the second result in the theorem. By (23), (26), (27) and (28) , for $k = 1, \cdots, K-2$, we have

$$
n^{-2} I_{kn} \to \sum_{l=K-1}^{K} 0.5\pi_{k,K}^b \pi_{l,K}^b \left(\frac{\Gamma_{kl,K}^{0b}[\Gamma_{K-1\cdot,K}^{0b} + \Gamma_{K\cdot,K}^{0b}]}{\Gamma_{l\cdot,K}^{0b}[\Gamma_{kK-1,K}^{b} + \Gamma_{kK,K}^{0b}]} - 1 \right)^2.
$$

Similarly, by (24), (26), (27) and (28), we have

$$
\begin{aligned}
& n^{-2} II_n \\
\to & 0.5(\pi_{K-1,K}^b)^2 \left(\frac{\Gamma_{K-1K-1,K}^{0b}[\Gamma_{K-1\cdot,K}^{0b} + \Gamma_{K\cdot,K}^{0b}]^2}{[\Gamma_{K-1\cdot,K}^{0b}]^2[\Gamma_{K-1K-1,K}^{0b} + 2\Gamma_{K-1K,K}^{0b} + \Gamma_{KK,K}^{0b}]} - 1 \right)^2 \\
& + \pi_{K-1,K}^b \pi_{K,K}^b \\
& \times \left(\frac{\Gamma_{K-1K,K}^{0b}[\Gamma_{K-1\cdot,K}^{0b} + \Gamma_{K\cdot,K}^{0b}]^2}{\Gamma_{K-1\cdot,K}^{0b}\Gamma_{K\cdot,K}^{0b}[\Gamma_{K-1K-1,K}^{0b} + 2\Gamma_{K-1K,K}^{0b} + \Gamma_{KK,K}^{0b}]} - 1 \right)^2 \\
& + 0.5(\pi_{K,K}^b)^2 \left(\frac{\Gamma_{KK,K}^{0b}[\Gamma_{K-1\cdot,K}^{0b} + \Gamma_{K\cdot,K}^{0b}]^2}{[\Gamma_{K\cdot,K}^{0b}]^2[\Gamma_{K-1K-1,K}^{0b} + 2\Gamma_{K-1K,K}^{0b} + \Gamma_{KK,K}^{0b}]} - 1 \right)^2.
\end{aligned}
$$

Clearly, there exits $c_{K2} < \infty$ such that

$$
n^{-2}\tilde{\mathcal{B}}_{K,n} = \sum_{k=1}^{K-2} n^{-2} I_{kn} + n^{-2} II_n \le c_{K2}.
$$

In addition, Assumption 5 implies that at least one of the squares is nonzero. Therefore, there exists a constant $c_{k1} > 0$ such that

$$
n^{-2}\tilde{\mathcal{B}}_{K,n} = \sum_{k=1}^{K-2} n^{-2} I_{kn} + n^{-2} II_n \ge c_{K1}.
$$

D.4 Proof of Theorem 6

We consider the upper bound for $L_n(\hat{Z}^b_{K_0+1}, \hat{Z}_{K_0})$. We say z is a $n \times (K_0+1)$ membership matrix for n nodes and K_0+1 groups if there is only one element in each row of z that takes value 1, and the rest of the entries are zero. Say $Z_{ik} = 1$, then we say that the i-th node is identified in group k. Let

$$\mathcal{V}_{K_0+1} = \left\{ \begin{array}{c} z \text{ is a } n \times (K_0+1) \text{ membership matrix s.t.} \\ \text{every group identified by z is a subset of} \\ \text{one of the true communities and} \\ \inf_{1 \le k \le K} n_k(z)/n \ge \varepsilon \end{array} \right\}.$$

Without loss of generality, we assume that $\hat{Z}^b_{K_0+1}$ is obtained by splitting the last group in $\hat{Z}_{K_0}$ into the K_0 -th and (K_0+1)-th groups in $\hat{Z}^b_{K_0+1}$. By Theorem 4 and Assumption 6, we have $\hat{Z}^b_{K_0+1} \in \mathcal{V}_{K_0+1}$ $a.s.$ Let z_{K_0+1} be an arbitrary realization of $\hat{Z}^b_{K_0+1}$ such that $z_{K_0+1} \in \mathcal{V}_{K_0+1}$ and $h(\cdot|z_{K_0+1})$ be a surjective mapping: $[K_0+1] \mapsto [K_0]$ that maps the community index identified by z_{K_0+1} into the true community index in $[K_0]$ for any $z_{K_0+1} \in \mathcal{V}_{K_0+1}$. Then, we have

$$h(k|z_{K_0+1}) = k, \quad k = 1, \cdots, K_0 - 1$$

and

$$h(K_0|z_{K_0+1}) = h(K_0+1|z_{K_0+1}) = K_0.$$

In the following, we explicitly write down the terms M_{kl}, $\widehat{M}_{kl}$, and O_{kl} as functions of z_{K_0+1}, i.e.,

$$M_{kl}(z_{K_0+1}) = \frac{E[O_{kl}(z_{K_0+1})]}{\sum_{i' \in \mathcal{C}_k(z_{K_0+1}), j' \in \mathcal{C}_l(z_{K_0+1}), i' \ne j'} d_{i'} d_{j'}}, \tag{39}$$

$$\widehat{M}_{kl}(z_{K_0+1}) = \frac{O_{kl}(z_{K_0+1})}{(\sum_{l'=1}^K O_{kl'}(z_{K_0+1}))(\sum_{l'=1}^K O_{ll'}(z_{K_0+1}))},$$

and

$$O_{kl}(z_{K_0+1}) = \sum_{i=1}^n \sum_{j \ne i} 1\{[z_{K_0+1}]_{ik} = 1, [z_{K_0+1}]_{jl} = 1\} A_{ij},$$

where $\mathcal{C}_l(z_{K_0+1})$ denotes the l-th cluster identified by z_{K_0+1}. Further recall n_{kl} and n_k defined in (6) in Section 3.3. We emphasize the dependence on z_{K_0+1} because, by Theorem 4, Z_K and Z^b_K for $K = 1, \cdots, K_0$ are uniquely defined, while $Z^b_{K_0+1}$ is not. By (39), for any $z_{K_0+1} \in \mathcal{V}_{K_0+1}$, $i \in \mathcal{C}_k(z_{K_0+1})$ and $j \in \mathcal{C}_l(z_{K_0+1})$, $k = 1, \cdots, K_0 - 1$, $l = K_0, K_0+1$. Then,

$$P_{ij}(z_{K_0+1}) = B_{h(k|z_{K_0+1})h(l|z_{K_0+1})} \theta_i \theta_j = B_{kK_0,K_0} \theta_i \theta_j = P_{ij}(Z_{K_0})$$

and

$$\frac{M_{kl}(z_{K_0+1})}{M_{kK_0,K_0}} = \frac{P_{ij}(z_{K_0+1})}{P_{ij}(Z_{K_0})} = 1, \quad k = 1, \cdots, K_0 - 1, \quad l = K_0, K_0+1. \tag{40}$$

Similarly,

$$\frac{M_{K_0K_0}(z_{K_0+1})}{M_{K_0K_0,K_0}} = \frac{M_{K_0K_0+1}(z_{K_0+1})}{M_{K_0K_0,K_0}} = \frac{M_{K_0+1K_0+1}(z_{K_0+1})}{M_{K_0K_0,K_0}} = 1. \tag{41}$$

By Theorem 4, $\hat{Z}_{K_0} = Z_{K_0}$ and $\hat{Z}^b_{K_0+1} \in \mathcal{V}_{K_0+1}$ $a.s.$ Therefore, (40) and (41) still hold when z_{K_0+1} and Z_{K_0} are replaced by $\hat{Z}^b_{K_0+1}$ and $\hat{Z}_{K_0}$. Then,

$$\begin{aligned}
&L_n(\hat{Z}^b_{K_0+1}, \hat{Z}_{K_0})\\
=&2\sum_{k=1}^{K_0-1}\sum_{l=K_0}^{K_0+1} 0.5 n_{kl}(\hat{Z}^b_{K_0+1})\Big(\frac{\widehat{M}_{kl}(\hat{Z}^b_{K_0+1})}{\widehat{M}_{kK_0,K_0}} - 1\Big)^2\\
&+ 0.5\Big[n_{K_0K_0}(\hat{Z}^b_{K_0+1})\Big(\frac{\widehat{M}_{K_0K_0}(\hat{Z}^b_{K_0+1})}{\widehat{M}_{K_0K_0,K_0}} - 1\Big)^2\\
&+ 2n_{K_0K_0+1}(\hat{Z}^b_{K_0+1})\Big(\frac{\widehat{M}_{K_0K_0+1}(\hat{Z}^b_{K_0+1})}{\widehat{M}_{K_0K_0,K_0}} - 1\Big)^2\\
&+ n_{K_0+1K_0+1}(\hat{Z}^b_{K_0+1})\Big(\frac{\widehat{M}_{K_0+1,K_0+1}(\hat{Z}^b_{K_0+1})}{\widehat{M}_{K_0K_0,K_0}} - 1\Big)^2\Big]. \qquad (42)
\end{aligned}$$

For the first term in (42),

$$0.5 n_{kl}(\hat{Z}^b_{K_0+1})\Big(\frac{\widehat{M}_{kl}(\hat{Z}^b_{K_0+1})}{\widehat{M}_{kK_0,K_0}} - 1\Big)^2 \lesssim n^2 \sup_{z_{K_0+1}\in\mathcal{V}_{K_0+1}}\Big(\frac{\widehat{M}_{kl}(z_{K_0+1})}{\widehat{M}_{kK_0,K_0}} - \frac{M_{kl}(z_{K_0+1})}{M_{kK_0,K_0}}\Big)^2.$$

The rate of the RHS of the above display depends on that of

$$\sup_{z_{K_0+1}\in\mathcal{V}_{K_0+1}} |O_{kl}(z_{K_0+1}) - E[O_{kl}(z_{K_0+1})]|.$$

By Bernstein inequality,

$$\begin{aligned}
&P(\sup_{z_{K_0+1}\in\mathcal{V}_{K_0+1}} |O_{kl}(z_{K_0+1}) - E[O_{kl}(z_{K_0+1})]| \geq Cn^{3/2}\rho_n^{1/2})\\
\leq& 2^n \exp\Big(-\frac{C^2 n^3\rho_n/2}{\overline{\theta}^2 n^2\rho_n + Cn^{3/2}\rho_n^{1/2}/3}\Big) \leq \exp(-C'n)
\end{aligned}$$

for some constant $C' > 0$. Therefore,

$$\sup_{z_{K_0+1}\in\mathcal{V}_{K_0+1}} |O_{kl}(z_{K_0+1}) - E[O_{kl}(z_{K_0+1})]| = O_{a.s.}(n^{3/2}\rho_n^{1/2}).$$

It also implies the uniform consistency that

$$\sup_{z_{K_0+1}\in\mathcal{V}_{K_0+1}} |n^{-2}\rho_n^{-1} O_{kl}(z_{K_0+1}) - \Gamma_{kl}(z_{K_0+1})| = O_{a.s.}((n\rho_n)^{-1/2}) + o(1) = o_{a.s.}(1),$$

where

$$\Gamma_{kl}(z_{K_0+1}) = \frac{n_l(z_{K_0+1})}{n}\frac{n_k(z_{K_0+1})}{n} H_{h(k|z_{K_0+1})h(l|z_{K_0+1})}.$$

Following the same and tedious Taylor expansion detailed in Steps 1 and 2 of the proof of Theorem 5, we have

$$\sup_{z_{K_0+1}\in\mathcal{V}_{K_0+1}} \Big|\widehat{M}_{kl}(z_{K_0+1}) - M_{kl}(z_{K_0+1})\Big| = O_{a.s.}((n^{5/2}\rho_n^{3/2})^{-1}),$$

$$|\widehat{M}_{kK_0,K_0} - M_{kK_0,K_0}| = O_p((n^3\rho_n^{3/2})^{-1}),$$

and

$$n^2\rho_n M_{kK_0,K_0} \geq c,$$

for some constant $c > 0$. Therefore,

$$\sup_{z_{K_0+1}\in\mathcal{V}_{K_0+1}} \left| \frac{\widehat{M}_{kl}(z_{K_0+1})}{\widehat{M}_{kK_0,K_0}} - \frac{M_{kl}(z_{K_0+1})}{M_{kK_0,K_0}} \right| = O_p((n\rho_n)^{-1/2})$$

and

$$0.5n_{kl}(\hat{Z}^b_{K_0+1})\left(\frac{\widehat{M}_{kl}(\hat{Z}^b_{K_0+1})}{\widehat{M}_{kK_0,K_0}} - 1\right)^2 = O_p(n\rho_n^{-1}).$$

The rest of the terms in (42) can be bounded similarly. Thus, we conclude that

$$L_n(\hat{Z}^b_{K_0+1}, \hat{Z}_{K_0}) = O_p(n\rho_n^{-1}). \tag{43}$$

Next, we study the asymptotic property of $\hat{K}_1$. If $K_0 = 1$, $P(\hat{K}_1 \geq 1) = 1$ holds trivially. If $K_0 \geq 2$,

$$R(1) \asymp \frac{n^2}{\eta_n} \asymp 1.$$

When $2 \leq K < K_0$, by Theorem 5,

$$R(K) \asymp \frac{\tilde{\mathcal{B}}_{K-1} + O_p(n\rho_n^{-1/2})}{\tilde{\mathcal{B}}_K + O_p(n\rho_n^{-1/2})} \asymp 1.$$

When $K = K_0$, by Theorem 5 and (43),

$$R(K_0) \lesssim \frac{n\rho_n^{-1}}{c_{K1}n^2 + O_p(n\rho_n^{-1/2})} \to 0.$$

Since $n^2/(n\rho_n^{-1}) = n\rho_n \geq C_1\log(n) \to \infty$ under Assumption 4,

$$P(\hat{K}_1 \geq K_0) \leq P\left(R(K_0) < \max_{K<K_0} R(K)\right) \to 1.$$

Now, we study the asymptotic property of $\tilde{K}_2$. If $K_0 = 1$,

$$R(1) \lesssim \frac{1}{n\rho_n} \to 0.$$

Therefore, $P(\tilde{K}_2 = 1) = P(R(1) \leq h_n) \to 1$ because $n\rho_n h_n \to \infty$ as $n \to \infty$. If $K_0 \geq 2$, by Theorem 5 and (43),

$$\begin{cases} R(K) \asymp & \frac{n^2}{n\rho_n} \to \infty, \quad \text{if } K = 1, \\ R(K) \asymp & 1, \quad \text{if } 2 \leq K < K_0, \\ R(K) \lesssim & \frac{n\rho_n^{-1}}{n^2} \asymp \frac{1}{n\rho_n} \to 0, \quad \text{if } K = K_0. \end{cases}$$

This, in conjunction with the conditions that $n\rho_n h_n \to \infty$ and $h_n \to 0$ as $n \to \infty$ implies that

$$P(\tilde{K}_2 = K_0) = P\left(\min_{1\leq K<K_0} R(K) > h_n, R(K_0) \leq h_n\right) \to 1.$$

It follows that $P(\hat{K}_2 = K_0) \geq P(\hat{K}_1 \geq K_0, \tilde{K}_2 = K_0) \to 1$.

Appendix E. Technical lemmas

Lemma 7 *Suppose Assumptions 1 and 2 hold. Let u_i^T be the i-th row of U_{1n}.*

(1) There exists a $K_0 \times K_0$ matrix S_n^τ such that $(S_n^\tau)^T S_n^\tau = I_{K_0}$ and $U_{1n} = \Theta_\tau^{1/2} Z_{K_0}(Z_{K_0}^T \Theta_\tau Z_{K_0})^{-1/2} S_n^\tau$.

(2) Let $[S_n^\tau](K)$ and $[S_n^\tau]_k(K)$ denote the first K columns of S_n^τ and its k-th row, respectively. There exist some $K \times K$ orthonormal matrix O_s, a $K_0 \times K_0$ matrix S_∞ and a positive constant c such that for any $K \leq K_0$, $[S_n^\tau]_k(K) O_s \to [S_\infty](K)$, $[S_\infty](K)$ has rank K, and for any $k = 1, \cdots, K_0$ and $K = 1, \cdots, K_0$,

$$\liminf_n ||[S_n^\tau]_k(K)|| \geq \underline{c}.$$

Proof The first result is proved in **?**. For part (2), by the proof of Theorem 1(2), we have

$$S_n^\tau[K] O_s \to S_\infty[K]$$

where S_∞ is the eigenvector matrix of $\Pi_\infty'^{1/2} H^*_{0,K_0} \Pi_\infty'^{1/2}$ and is of full rank, and O_s is a $K \times K$ orthogonal matrix. In addition, by Assumptions 1(2) and 2, all elements in $\Pi_\infty'^{1/2} H^*_{0,K_0} \Pi_\infty'^{1/2}$ are positive. By **?**, Lemma 8.2.1, all elements in the first column of S_∞ are strictly positive. This implies that, for any $k = 1, \cdots, K_0$,

$$\liminf_n ||[S_n^\tau]_k(K)|| = \liminf_n ||[S_n^\tau]_k(K) O_s|| = ||[S_\infty]_k(K)|| \geq ||[S_\infty]_{k1}|| > 0.$$

This concludes the proof. ■

The following lemma is largely based on **?**, Theorem 3.2 and **?**, Theorem 2.3.

Lemma 8 *Let $\mathcal{C}$ be a set of nodes and $\{\hat{\beta}_{in}\}_{i \in \mathcal{C}}$ be a sequence of $d_\beta \times 1$ vectors such that $\sup_{i \in \mathcal{C}} ||\hat{\beta}_{in} - \beta_{in}|| \leq c_1$ a.s. and $\sup_{i \in \mathcal{C}} ||\beta_{in}|| \leq M$ for some sufficiently small constant $c_1 > 0$ and some constant $M > 0$, respectively. In addition, suppose $\{\beta_{in}\}_{i \in \mathcal{C}}$ has L distinct vectors for some $L \geq K$ and we group index i into L mutually exclusive groups $\{\mathcal{C}_l\}_{l=1}^L$ such that if $i, j \in \mathcal{C}_l$, $\beta_{in} = \beta_{jn}$ and for any $i \in \mathcal{C}_l$, $j \in \mathcal{C}_{l'}$, $l \neq l'$, $\inf_{i,j,n} ||\beta_{in} - \beta_{jn}|| > c_2 > 0$. Let $\pi_l = \frac{\#\mathcal{C}_l}{n}$, $l = 1, \cdots, L$. Then, $\min_{l=1,\cdots,L} \pi_l \geq \underline{\pi} > 0$. We apply k-means algorithm on $\{\beta_{in}\}_{i=1}^n$ and $\{\hat{\beta}_{in}\}_{i=1}^n$ and obtain K sets of mutually exclusive groups $(\mathcal{C}(1), \cdots, \mathcal{C}(K))$ and $(\widehat{\mathcal{C}}(1), \cdots, \widehat{\mathcal{C}}(K))$, respectively. Suppose $\mathcal{C}(k)$, $k = 1, \cdots, K$ are uniquely defined, then*

(1) for any $l = 1, \cdots, L$,

$$\mathcal{C}_l \subset \text{one of } \{\mathcal{C}(k), k = 1, \cdots, K\};$$

(2)

$$\left| \frac{\widehat{\Phi}(\mathcal{C}) - \sum_{k=1}^K \widehat{\Phi}(\widehat{\mathcal{C}}(k))}{\#\mathcal{C}} - \frac{\Phi(\mathcal{C}) - \sum_{k=1}^k \Phi(\mathcal{C}(k))}{\#\mathcal{C}} \right| \leq C c_1, \ a.s.,$$

where $C > 0$ is some constant independent of n and for a generic index set $\mathcal{C}$,

$$\widehat{\Phi}(\mathcal{C}) = \sum_{i \in \mathcal{C}} ||\hat{\beta}_{in} - \frac{\sum_{i \in \mathcal{C}} \hat{\beta}_{in}}{\#\mathcal{C}}||^2$$

and

$$\Phi(\mathcal{C}) = \sum_{i \in \mathcal{C}} ||\beta_{in} - \frac{\sum_{i \in \mathcal{C}} \beta_{in}}{\#\mathcal{C}}||^2; \quad \textit{and}$$

(3) after relabeling, $\widehat{\mathcal{C}}(k) = \mathcal{C}(k)$, $k = 1, \cdots, K$.

Proof Following the proof of **?**, Theorem 3.2, we focus on the case $L = 3$. The proof for $L \geq 4$ is similar but require more notation. When $K = 1$, the results are trivial. When $K = 3$, Lemma 8(1) is trivial as $\mathcal{C}(k) = \mathcal{C}_k$, $k = 1, 2, 3$ after relabeling. Lemma 8(3) directly follows **?**, Theorem 2.3, given that c_1 is sufficiently small so that

$$(2c_1 \underline{\pi}^{1/2} + 16K^{3/4}M^{1/2}c_1)^2 \leq \underline{\pi} c_2^2.$$

Given Lemma 8(3), Lemma 8(2) holds with $C = 16M$ because

$$\left| \left\| \hat{\beta}_{in} - \frac{\sum_{i \in \mathcal{C}} \hat{\beta}_{in}}{\#\mathcal{C}} \right\|^2 - \left\| \beta_{in} - \frac{\sum_{i \in \mathcal{C}} \beta_{in}}{\#\mathcal{C}} \right\|^2 \right| \leq 8Mc_1.$$

Next, we proof Lemma 8 for $K = 2$. Denote $\bar{\beta}_l$, $l = 1, 2, 3$ as the true values β_{in} can take when $i \in \mathcal{C}_1, \mathcal{C}_2$, and $\mathcal{C}_3$, respectively.

Step 1. Proof of Lemma 8(1). Suppose

$$\frac{\pi_2 \pi_3}{\pi_2 + \pi_3} ||\bar{\beta}_2 - \bar{\beta}_3||^2 < \frac{\pi_1 \pi_3}{\pi_1 + \pi_3} ||\bar{\beta}_1 - \bar{\beta}_3||^2 < \frac{\pi_1 \pi_2}{\pi_1 + \pi_2} ||\bar{\beta}_1 - \bar{\beta}_2||^2 \tag{44}$$

In this case, we aim to show that $\mathcal{C}(1) = \mathcal{C}_1$ and $\mathcal{C}(2) = \mathcal{C}_2 \cup \mathcal{C}_3$. Suppose that, by the k-means algorithm, $n\pi_l^*$ nodes of $i \in \mathcal{C}_l$, $\pi_l^* \in [0, \pi_l]$, $l = 1, 2, 3$ are classified into $\mathcal{C}(1)$ and the rest are in $\mathcal{C}(2)$. We aim to show that (44) implies $\pi_1^* = \pi_1$ and $\pi_2^* = \pi_3^* = 0$. The k-means objective function for the classification $(\mathcal{C}(1), \mathcal{C}(2))$ is

$$F(\alpha_1, \alpha_2; \pi_1^*, \pi_2^*, \pi_3^*) \equiv \sum_{l=1}^{3} \pi_l^* ||\bar{\beta}_l - \alpha_1||^2 + \sum_{l=1}^{3} (\pi_l - \pi_l^*) ||\bar{\beta}_l - \alpha_2||^2,$$

where $\alpha_1 = \frac{\sum_{l=1}^3 \pi_l^* \bar{\beta}_l}{\sum_{l=1}^3 \pi_l^*}$ and $\alpha_2 = \frac{\sum_{l=1}^3 (\pi - \pi_l^*) \bar{\beta}_l}{\sum_{l=1}^3 (\pi - \pi_l^*)}$. Suppose $\pi_1^* \in (0, \pi_1)$, then we have

$$||\bar{\beta}_1 - \alpha_1|| = ||\bar{\beta}_1 - \alpha_2||,$$

which implies that, for any $\tilde{\pi} \in (0, \pi)$,

$$F(\alpha_1, \alpha_2; \pi_1^*, \pi_2^*, \pi_3^*) = F(\alpha_1, \alpha_2; \tilde{\pi}, \pi_2^*, \pi_3^*) \geq F(\tilde{\alpha}_1, \tilde{\alpha}_2; \tilde{\pi}, \pi_2^*, \pi_3^*),$$

where $\tilde{\alpha}_1 = \frac{\tilde{\pi}_1 \bar{\beta}_1 + \pi_2^* \bar{\beta}_2 + \pi_3^* \bar{\beta}_3}{\tilde{\pi}_1 + \pi_2^* + \pi_3^*}$ and $\tilde{\alpha}_2 = \frac{(\pi_1 - \tilde{\pi}_1)\bar{\beta}_1 + (\pi_2 - \pi_2^*)\bar{\beta}_2 + (\pi_3 - \pi_3^*)\bar{\beta}_3}{1 - \tilde{\pi}_1 - \pi_2^* - \pi_3^*}$ are the minimizer of $F(\cdot, \cdot; \tilde{\pi}, \pi_2^*, \pi_3^*)$. In addition, because $F(\alpha_1, \alpha_2; \pi_1^*, \pi_2^*, \pi_3^*)$ achieves the minimum of the k-means objective function among all classifications, we have

$$F(\alpha_1, \alpha_2; \pi_1^*, \pi_2^*, \pi_3^*) \leq F(\tilde{\alpha}_1, \tilde{\alpha}_2; \tilde{\pi}, \pi_2^*, \pi_3^*),$$

which implies that the equality holds, for any $\tilde{\pi}_1 \in (0, \pi_1)$. Then, by the uniqueness of the minimizer for the quadratic objective function $F(\cdot, \cdot; \tilde{\pi}, \pi_2^*, \pi_3^*)$, we have, for any $\tilde{\pi} \in (0, \pi_1)$,

$$(\alpha_1, \alpha_2) = (\tilde{\alpha}_1, \tilde{\alpha}_2).$$

This implies that $\bar{\beta}_1 = \frac{\pi_2^* \bar{\beta}_2 + \pi_3^* \bar{\beta}_3}{\pi_2^* + \pi_3^*} = \frac{(\pi_2 - \pi_2^*)\bar{\beta}_2 + (\pi_3 - \pi_3^*)\bar{\beta}_3}{\pi_2 - \pi_2^* + (\pi_3 - \pi_3^*)} = \frac{\pi_2 \bar{\beta}_2 + \pi_3 \bar{\beta}_3}{\pi_2 + \pi_3}$. Plugging this equality into (44), we have

$$\frac{\pi_2 \pi_3}{\pi_2 + \pi_3} ||\bar{\beta}_2 - \bar{\beta}_3||^2 < \frac{\pi_1 \pi_2}{\pi_1 + \pi_2} ||\bar{\beta}_1 - \bar{\beta}_2||^2 = \left(\frac{\pi_1}{\pi_1 + \pi_2}\right)\left(\frac{\pi_3}{\pi_2 + \pi_3}\right)\left(\frac{\pi_2 \pi_3}{\pi_2 + \pi_3} ||\bar{\beta}_2 - \bar{\beta}_3||^2\right),$$

which is a contradiction. This implies that $\pi_1^* = 0$ or π_1. Similarly, if $\pi_2^* \in (0, \pi_2)$, we can show that $\bar{\beta}_2 = \frac{\pi_1 \bar{\beta}_1 + \pi_3 \bar{\beta}_3}{\pi_1 + \pi_3}$. Then, by (44),

$$\frac{\pi_1 \pi_3}{\pi_1 + \pi_3} ||\bar{\beta}_1 - \bar{\beta}_3||^2 < \frac{\pi_1 \pi_2}{\pi_1 + \pi_2} ||\bar{\beta}_1 - \bar{\beta}_2||^2 = \left(\frac{\pi_3}{\pi_1 + \pi_2}\right)\left(\frac{\pi_2}{\pi_2 + \pi_3}\right)\left(\frac{\pi_1 \pi_3}{\pi_1 + \pi_3} ||\bar{\beta}_1 - \bar{\beta}_3||^2\right),$$

which is again a contradiction. Therefore, $\pi_2^* = 0$ or π_2. This means, $\mathcal{C}_k \subset \mathcal{C}(1)$ or $\mathcal{C}(2)$, for $k = 1, 2$. Last, we assume the k-means algorithm classify π_3^* fraction of $\mathcal{C}_3$ with $\mathcal{C}_1$ and the rest with $\mathcal{C}_2$. Then, the k-means objective function becomes

$$\min_{\alpha_1, \alpha_2} F(\alpha_1, \alpha_2; \pi_1, \pi_2, \pi_3^*) = \frac{\pi_1 \pi_3^*}{\pi_1 + \pi_3^*} ||\bar{\beta}_1 - \bar{\beta}_3||^2 + \frac{\pi_2 (\pi_3 - \pi_3^*)}{\pi_2 + \pi_3 - \pi_3^*} ||\bar{\beta}_2 - \bar{\beta}_3||^2.$$

When $\pi_3^* = 0$, the above display becomes $\frac{\pi_2 \pi_3}{\pi_2 + \pi_3} ||\bar{\beta}_2 - \bar{\beta}_3||^2$. In addition,

$$\begin{aligned}
&\left(\frac{\pi_1 \pi_3^*}{\pi_1 + \pi_3^*} ||\bar{\beta}_1 - \bar{\beta}_3||^2 + \frac{\pi_2 (\pi_3 - \pi_3^*)}{\pi_2 + \pi_3 - \pi_3^*} ||\bar{\beta}_2 - \bar{\beta}_3||^2\right) - \left(\frac{\pi_2 \pi_3}{\pi_2 + \pi_3} ||\bar{\beta}_2 - \bar{\beta}_3||^2\right) \\
=&\pi_3^* \left(\frac{\pi_1}{\pi_1 + \pi_3^*} ||\bar{\beta}_1 - \bar{\beta}_3||^2 - \frac{\pi_2^2}{(\pi_2 + \pi_3)(\pi_2 + \pi_3 - \pi_3^*)} ||\bar{\beta}_2 - \bar{\beta}_3||^2\right) \\
\geq&\pi_3^* \left(\frac{\pi_1}{\pi_1 + \pi_3} ||\bar{\beta}_1 - \bar{\beta}_3||^2 - \frac{\pi_2}{(\pi_2 + \pi_3)} ||\bar{\beta}_2 - \bar{\beta}_3||^2\right) \geq 0,
\end{aligned}$$

where the first inequality holds because the term in the parenthesis after the first equal sign is a decreasing function in $\pi_3^* \in [0, \pi_3]$ and the last inequality holds because of (44). This implies that $\pi_3^* = 0$, i.e., $\mathcal{C}(1) = \mathcal{C}_1$ and $\mathcal{C}(2) = \mathcal{C}_2 \cup \mathcal{C}_3$, which implies Lemma 8(1).

If the three terms in (44) take distinctive values, the above argument is valid after relabeling. If at least two terms take same values, then the k-means algorithm applying to $\{\beta_{in}\}_{i=1}^n$ do not have a unique solution. This situation has been ruled out by our assumption.

Step 2. Proof of Lemma 8(3). Let $\mathcal{Q}_n(\mathcal{A}) = \sum_{l=1}^L \min_{1 \leq k \leq K} \|\bar{\beta}_l - \alpha_k\|^2 \pi_k$, $\mathcal{A} \in \mathcal{M} = \{(\alpha_1, \ldots, \alpha_K) : \sup_{1 \leq k \leq K} \|\alpha_k\| \leq 2M\}$ for some constant M independent of n, $g_i^0 = k$ if $i \in \mathcal{C}(k)$, and $R_n = \sup_i \|\hat{\beta}_{in} - \beta_{in}\|$. By the assumptions in Lemma 8,

$$R_n \leq c_1 \quad a.s. \tag{45}$$

In addition,

$$\begin{aligned}\|\hat{\beta}_{in}-\alpha_k\|^2 &\geq \|\beta_{in}-\alpha_k\|^2 - 2|(\beta_{in}-\hat{\beta}_{in})^T(\beta_{in}-\alpha_l)| - \|\beta_{in}-\hat{\beta}_{in}\|^2\\ &\geq \|\beta_{in}-\alpha_k\|^2 - 2\|\beta_{in}-\hat{\beta}_{in}\|\|\beta_{in}-\alpha_k\| - R_n^2\\ &\geq \|\beta_{in}-\alpha_k\|^2 - 6MR_n - R_n^2\\ &\geq \|\beta_{in}-\alpha_k\|^2 - 7MR_n,\end{aligned}$$

where the third inequality follows the Cauchy-Schwarz inequality. Taking $\min_{1\leq k\leq K}$ on both sides and averaging over i, we have

$$\begin{aligned}\widehat{\mathcal{Q}}_n(\mathcal{A}) \equiv & n^{-1}\sum_{i=1}^n \min_{1\leq k\leq K} ||\hat{\beta}_{in}-\alpha_l||^2\\ \geq & n^{-1}\sum_{i=1}^n \min_{1\leq k\leq K} ||\beta_{in}-\alpha_l||^2 - 7MR_n \geq \mathcal{Q}_n(\mathcal{A}) - 7Mc_1,\end{aligned}$$

where the inequality is due to (45). Similarly, we have $\widehat{\mathcal{Q}}_n(\mathcal{A}) \leq \mathcal{Q}_n(\mathcal{A}) + 7Mc_1$, and thus,

$$\breve{R}_n \equiv \sup_{\mathcal{A}\in\mathcal{M}} |\widehat{\mathcal{Q}}_n(\mathcal{A}) - \mathcal{Q}_n(\mathcal{A})| \leq 7Mc_1 \quad a.s. \tag{46}$$

We maintain (44). In this case, the minimizer of $\mathcal{Q}_n(\cdot)$, as shown in the previous step, is $\mathcal{A}^* = (\alpha_1^*, \alpha_2^*)$, where $\alpha_1^* = \bar{\beta}_1$ and $\alpha_2^* = \frac{\pi_2\bar{\beta}_2+\pi_3\bar{\beta}_3}{\pi_2+\pi_3}$. Then, $Q_n(\mathcal{A}^*) = \frac{\pi_2\pi_3}{\pi_2+\pi_3}||\bar{\beta}_2-\bar{\beta}_3||^2$. For a generic $\mathcal{A} = (\alpha_1, \alpha_2)$ and $\mathcal{H}(\mathcal{A}, \mathcal{A}^*) \geq \eta$, where $\mathcal{H}(\cdot,\cdot)$ denotes the Hausdorff distance of two sets, we aim to lower bound $\mathcal{Q}_n(\mathcal{A}) - \mathcal{Q}_n(\mathcal{A}^*)$. In view of the definition of $Q_n(\cdot)$, we consider the following three cases: between α_1 and α_2,

(1) $\bar{\beta}_1$ is closer to α_1 while $(\bar{\beta}_2, \bar{\beta}_3)$ are closer to α_2;

(2) $\bar{\beta}_2$ is closer to one of α_1 while $(\bar{\beta}_1, \bar{\beta}_3)$ are closer to α_2;

(3) $\bar{\beta}_3$ is closer to one of α_1 while $(\bar{\beta}_1, \bar{\beta}_2)$ are closer to α_2;

(4) $(\bar{\beta}_1, \bar{\beta}_2, \beta_3)$ are all closer to one of α_1 and α_2.

For case (1),

$$\begin{aligned}\mathcal{Q}_n(\mathcal{A}) - \mathcal{Q}_n(\mathcal{A}^*) =& \pi_1||\bar{\beta}_1-\alpha_1||^2 + \sum_{l=2,3}\pi_l\left[||\bar{\beta}_l-\alpha_2||^2 - ||\bar{\beta}_l-\alpha_2^*||^2\right]\\ =& \pi_1||\alpha_1^*-\alpha_1||^2 + \sum_{l=2,3}\pi_l\left[2(\bar{\beta}_l-\alpha_2^*)^T(\alpha_2^*-\alpha_2) + ||\alpha_2-\alpha_2^*||^2\right]\\ =& \pi_1||\alpha_1^*-\alpha_1||^2 + (\pi_2+\pi_3)||\alpha_2-\alpha_2^*||^2\\ \geq& \underline{\pi}\max(||\alpha_1^*-\alpha_1||, ||\alpha_2-\alpha_2^*||)^2 \geq \underline{\pi}\eta^2,\end{aligned}$$

where the third equality holds because $\alpha_2^* = \frac{\pi_2\bar{\beta}_2+\pi_3\bar{\beta}_3}{\pi_2+\pi_3}$, the first inequality holds because for arbitrary constants $a, b > 0$, $a + b \geq \max(a, b)$, and the last inequality holds because,

$$\mathcal{H}(\mathcal{A}, \mathcal{A}^*) = \max(\mathcal{H}_1(\mathcal{A}, \mathcal{A}^*), \mathcal{H}_2(\mathcal{A}, \mathcal{A}^*)),$$

where

$$\begin{aligned}\mathcal{H}_1(\mathcal{A},\mathcal{A}^*) =&\max(\min(||\alpha_1^*-\alpha_1||,||\alpha_1^*-\alpha_2||),\min(||\alpha_2^*-\alpha_1||,||\alpha_2^*-\alpha_2||))\\ \leq&\max(||\alpha_1^*-\alpha_1||,||\alpha_2^*-\alpha_2||)\end{aligned}$$

and

$$\begin{aligned}\mathcal{H}_2(\mathcal{A},\mathcal{A}^*) =&\max(\min(||\alpha_1^*-\alpha_1||,||\alpha_1-\alpha_2^*||),\min(||\alpha_2-\alpha_1^*||,||\alpha_2^*-\alpha_2||))\\ \leq&\max(||\alpha_1^*-\alpha_1||,||\alpha_2^*-\alpha_2||).\end{aligned}$$

For case (2), we have

$$\begin{aligned}Q_n(\mathcal{A})-Q_n(\mathcal{A}^*)\geq&\inf_{\alpha_2}\left(\pi_1||\bar{\beta}_1-\alpha_2||^2+\pi_3||\bar{\beta}_3-\alpha_2||^2\right)-\frac{\pi_2\pi_3}{\pi_2+\pi_3}||\bar{\beta}_2-\bar{\beta}_3||^2\\ \geq&\frac{\pi_1\pi_3}{\pi_1+\pi_3}||\bar{\beta}_1-\bar{\beta}_3||^2-\frac{\pi_2\pi_3}{\pi_2+\pi_3}||\bar{\beta}_2-\bar{\beta}_3||^2\geq\underline{M}>0.\end{aligned}$$

where

$$\underline{M}=\min\left(\frac{\pi_1\pi_3}{\pi_1+\pi_3}||\bar{\beta}_1-\bar{\beta}_3||^2,\frac{\pi_1\pi_2}{\pi_1+\pi_2}||\bar{\beta}_1-\bar{\beta}_2||^2\right)-\frac{\pi_2\pi_3}{\pi_2+\pi_3}||\bar{\beta}_2-\bar{\beta}_3||^2$$

and the last inequality holds by (44).

Similarly, for case (3), we have

$$\begin{aligned}Q_n(\mathcal{A})-Q_n(\mathcal{A}^*)\geq&\inf_{\alpha_2}\left(\pi_1||\bar{\beta}_1-\alpha_2||^2+\pi_2||\bar{\beta}_2-\alpha_2||^2\right)-\frac{\pi_2\pi_3}{\pi_2+\pi_3}||\bar{\beta}_2-\bar{\beta}_3||^2\\ \geq&\frac{\pi_1\pi_2}{\pi_1+\pi_2}||\bar{\beta}_1-\bar{\beta}_2||^2-\frac{\pi_2\pi_3}{\pi_2+\pi_3}||\bar{\beta}_2-\bar{\beta}_3||^2\geq\underline{M}>0.\end{aligned}$$

Last, for the same reason, for case (4),

$$Q_n(\mathcal{A})-Q_n(\mathcal{A}^*)\geq\underline{M}>0. \tag{47}$$

Therefore, we have

$$\inf_{\mathcal{H}(\mathcal{A},\mathcal{A}^*)\geq\eta}Q_n(\mathcal{A})-Q_n(\mathcal{A}^*)\geq\min(\underline{\pi}\eta^2,\underline{M}).$$

Further define $\hat{\mathcal{A}}_n=(\hat{\alpha}_1,\hat{\alpha}_2)=\arg\min_{\mathcal{A}}\hat{Q}_n(\mathcal{A})$. Note $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are weighted average of $\{\hat{\beta}_{in}\}_{i=1}^n$ and $\sup_i||\hat{\beta}_{in}||\leq M+c_1\leq 2M$. Therefore, by (46),

$$|\hat{Q}_n(\hat{\mathcal{A}}_n)-Q_n(\hat{\mathcal{A}}_n)|\leq 7Mc_1,\quad a.s. \tag{48}$$

and

$$|\hat{Q}_n(\mathcal{A}^*)-Q_n(\mathcal{A}^*)|\leq 7Mc_1,\quad a.s. \tag{49}$$

Then,

$$\begin{aligned}&P(\mathcal{H}(\hat{\mathcal{A}}_n,\mathcal{A}^*)\geq(15M/\underline{\pi})^{1/2}c_1^{1/2}\quad i.o.)\\ =&P(\mathcal{H}(\hat{\mathcal{A}}_n,\mathcal{A}^*)\geq(15M/\underline{\pi})^{1/2}c_1^{1/2},\ Q_n(\hat{\mathcal{A}}_n)-Q_n(\mathcal{A}^*)\geq\min(15Mc_1,\underline{M})\quad i.o.)\\ \leq&P(14Mc_1+\hat{Q}_n(\hat{\mathcal{A}}_n)-\hat{Q}_n(\mathcal{A}^*)\geq\min(15Mc_1,\underline{M})\quad i.o.)\\ \leq&P(14Mc_1\geq\min(15Mc_1,\underline{M})\quad i.o.)\\ =&0,\end{aligned}$$

where the first equality holds due to (47), the first inequality holds because of (48) and (49), the second inequality holds because $\hat{Q}_n(\hat{\mathcal{A}}_n) - \hat{Q}_n(\mathcal{A}^*) \geq 0$, and the last equality holds because c_1 is sufficiently small so that $15Mc_1 \leq \underline{M}$. This implies

$$\mathcal{H}(\hat{\mathcal{A}}_n, \mathcal{A}^*) \leq (15M/\underline{\pi})^{1/2} c_1^{1/2}, \quad a.s.$$

Further note that $||\alpha_1^* - \alpha_2^*|| > 0$, otherwise $\bar{\beta}_1 = \frac{\pi_2\bar{\beta}_2 + \pi_3\bar{\beta}_3}{\pi_2+\pi_3}$ which is a contradiction as shown in Step 1. Let c_1 be sufficiently small so that $(15M/\underline{\pi})^{1/2} c_1^{1/2} < ||\alpha_1^* - \alpha_2^*||$. Then there is a one-to-one mapping $\mathcal{F}_n$: $\{1,2\} \mapsto \{1,2\}$ such that

$$\sup_{k=1,2} ||\hat{\alpha}_k - \alpha^*_{\mathcal{F}_n(k)}|| \leq (15M/\underline{\pi})^{1/2} c_1^{1/2}.$$

W.l.o.g., we assume $\mathcal{F}_n(k) = k$ such that

$$\sup_{k=1,2} ||\hat{\alpha}_k - \alpha^*_k|| \leq (15M/\underline{\pi})^{1/2} c_1^{1/2}.$$

Denote $\hat{g}_i = k$ if $i \in \widehat{\mathcal{C}}(k)$, $k = 1,2$ and $g_i^0 = k$ if $i \in \mathcal{C}(k)$, $k = 1,2$. If $\hat{g}_i \neq g_i^0$, then $||\hat{\beta}_{in} - \hat{\alpha}_{\hat{g}_i}|| \leq ||\hat{\beta}_{in} - \hat{\alpha}_{g_i^0}||$. Therefore,

$$\begin{aligned}
&||\beta_{in} - \alpha_{g_i^0}|| + c_1 + (15M/\underline{\pi})^{1/2} c_1^{1/2} \\
\geq & ||\hat{\beta}_{in} - \hat{\alpha}_{g_i^0}|| \\
\geq & ||\hat{\beta}_{in} - \hat{\alpha}_{\hat{g}_i}|| \geq ||\beta_{in} - \alpha^*_{\hat{g}_i}|| - c_1 - (15M/\underline{\pi})^{1/2} c_1^{1/2}.
\end{aligned}$$

Therefore,

$$1\{\hat{g}_i \neq g_i^0\} \leq 1\{2c_1 + 2(15M/\underline{\pi})^{1/2} c_1^{1/2} \geq ||\beta_{in} - \alpha^*_{\hat{g}_i}|| - ||\beta_{in} - \alpha^*_{g_i^0}||\} \quad a.s.$$

By Lemma 8(1), we only need to consider the lower bound for the RHS of the above display in three cases: (1) $g_i^0 = 1$ and $\beta_{in} = \bar{\beta}_1$, (2) $g_i^0 = 2$ and $\beta_{in} = \bar{\beta}_2$, and (3) $g_i^0 = 2$ and $\beta_{in} = \bar{\beta}_3$. For case (1),

$$||\beta_{in} - \alpha^*_{\hat{g}_i}|| - ||\beta_{in} - \alpha^*_{g_i^0}|| = ||\alpha_1^* - \alpha_2^*|| = \left\|\bar{\beta}_1 - \frac{\pi_2\bar{\beta}_2 + \pi_3\bar{\beta}_3}{\pi_2 + \pi_3}\right\| > 0,$$

where the last inequality holds because by the argument in Step 1, $\bar{\beta}_1 \neq \frac{\pi_2\bar{\beta}_2 + \pi_3\bar{\beta}_3}{\pi_2+\pi_3}$.

For case (2), $\alpha^*_{\hat{g}_i} = \alpha^*_1 = \bar{\beta}_1$ and

$$\begin{aligned}
||\beta_{in} - \alpha^*_{\hat{g}_i}|| - ||\beta_{in} - \alpha^*_{g_i^0}|| =& ||\bar{\beta}_2 - \bar{\beta}_1|| - \frac{\pi_3}{\pi_2 + \pi_3}||\bar{\beta}_2 - \bar{\beta}_3|| \\
\geq & ||\bar{\beta}_2 - \bar{\beta}_3|| \sqrt{\frac{\pi_3}{\pi_2 + \pi_3}} \left(\sqrt{\frac{\pi_1 + \pi_2}{\pi_1}} - \sqrt{\frac{\pi_3}{\pi_2 + \pi_3}}\right) > 0,
\end{aligned}$$

where the first inequality holds due to (44). Similarly, for case (3), we have

$$\begin{aligned}
||\beta_{in} - \alpha^*_{\hat{g}_i}|| - ||\beta_{in} - \alpha^*_{g_i^0}|| =& ||\bar{\beta}_3 - \bar{\beta}_1|| - \frac{\pi_2}{\pi_2 + \pi_3}||\bar{\beta}_2 - \bar{\beta}_3|| \\
\geq & ||\bar{\beta}_2 - \bar{\beta}_3|| \sqrt{\frac{\pi_2}{\pi_2 + \pi_3}} \left(\sqrt{\frac{\pi_1 + \pi_3}{\pi_1}} - \sqrt{\frac{\pi_2}{\pi_2 + \pi_3}}\right) > 0.
\end{aligned}$$

Let constant $\underline{C}$ be

$$\min\left(\left\|\bar{\beta}_1 - \frac{\pi_2\bar{\beta}_2 + \pi_3\bar{\beta}_3}{\pi_2 + \pi_3}\right\|, ||\bar{\beta}_2 - \bar{\beta}_1|| - \frac{\pi_3}{\pi_2 + \pi_3}||\bar{\beta}_2 - \bar{\beta}_3||, ||\bar{\beta}_3 - \bar{\beta}_1|| - \frac{\pi_2}{\pi_2 + \pi_3}||\bar{\beta}_2 - \bar{\beta}_3||\right) \geq \underline{C}$$

such that $\underline{C} > 0$. Then,

$$\begin{aligned} 1\{\hat{g}_i \neq g_i^0\} \leq & 1\{2c_1 + 2(15M/\underline{\pi})^{1/2}c_1^{1/2} \geq ||\beta_{in} - \alpha^*_{\hat{g}_i}|| - ||\beta_{in} - \alpha^*_{g_i^0}||\} \\ \leq & 1\{2c_1 + 2(15M/\underline{\pi})^{1/2}c_1^{1/2} \geq \underline{C}\}. \end{aligned}$$

Noting that the RHS of the above display is independent of i and choosing c_1 sufficiently small such that

$$2c_1 + 2(15M/\underline{\pi})^{1/2}c_1^{1/2} < \underline{C},$$

we have

$$P(\sup_i 1\{\hat{g}_i \neq g_i^0\} > 0,\ i.o.) \leq P(2c_1 + 2(15M/\underline{\pi})^{1/2}c_1^{1/2} \geq \underline{C},\ i.o.) = 0$$

This concludes that $\widehat{\mathcal{C}}(k) = \mathcal{C}(k)$ for $k = 1, 2$, which is the desired result for Lemma 8(3).

Step 3. Proof of Lemma 8(2). Given Lemma 8(3), the desired results can be derived by the same argument for $K = 3$. ■