On consistency of model selection for stochastic block models

Jianwei Hu* Hong Qin† Ting Yan‡ Yunpeng Zhao§
*,†,‡Central China Normal University and §George Mason University

Abstract

Estimating the number of communities is one of the fundamental problems in community detection. We re-examine the Bayesian paradigm for stochastic block models and propose a "corrected Bayesian information criterion", to determine the number of communities and show that the proposed estimator is consistent under mild conditions. The proposed criterion improves those used in Wang and Bickel (2016) and Saldana, Yu and Feng (2017) which tend to underestimate and overestimate the number of communities, respectively. Along the way, we establish the Wilks theorem for stochastic block models. Moreover, we show that, to obtain the consistency of model selection for stochastic block models, we need a so-called "consistency condition". We also provide sufficient conditions for both homogenous networks and non-homogenous networks. The results are further extended to degree corrected stochastic block models. Numerical studies demonstrate our theoretical results.

Key words: Consistency; Network data; Stochastic block model; SCORE; Spectral clustering

Mathematics Subject Classification: Primary 62E20; Secondary 60G05.

1 Introduction

Community structure within network data is quite common. For instance, peoples form groups in social networks based on common locations, interests, occupations and so on; proteins form communities based on functions in metabolic networks; papers form communities by research topics in citation networks. As a result, the links (or edges) among nodes are dense within communities and relatively sparse inter-communities. Identifying such sub-groups provides important insights into network formation mechanism and how network topology affects each other.

The stochastic block model (SBM) proposed by Holland, Laskey and Leinhardt (1983) is the best studied network models to detect community structure. For its applications, see Snijders

*Department of Statistics, Central China Normal University, Wuhan, 430079, China. Email: jwhu@mail.ccnu.edu.cn.

†Department of Statistics, Central China Normal University, Wuhan, 430079, China. Email: qinhong@mail.ccnu.edu.cn.

‡Department of Statistics, Central China Normal University, Wuhan, 430079, China. Email: tingyanty@mail.ccnu.edu.cn

§Department of statistics, George Mason University, Virginia 22030-4444, USA. Email: yzhao15@gmu.edu

and Nowicki (1997) and Nowicki and Snijders (2001). Let $A \in \{0,1\}^{n\times n}$ be the symmetric adjacency matrix of an undirected graph with n nodes. In the stochastic block model with k communities, each node is associated with one community labeled by $z_k(i)$, where $z_k(i) \in [k]$. Here $[m] = \{1,\ldots,m\}$ for any positive integer m. In other words, the nodes are given a community assignment $z_k : [n] \to [k]^n$. The diagonal entries of A are all zeros and the entries of the upper triangle matrix A are independent Bernoulli random variables with success probabilities $\{P_{ij}\}$ which only depend on the community labels of nodes i and j. That is, given the node communities, all edges are independent and for certain probability matrix $\theta_k = \{\theta_{kab}\}_{1\le a\le b\le k}$,

$$P(A_{ij} = 1 \mid z_k(i), z_k(j)) = \theta_{kz_k(i)z_k(j)}.$$

For simplicity, θ_k and z_k are abbreviated to θ and z, respectively.

A wide variety of methods have been proposed to estimate the latent block membership of nodes in an stochastic block model, including modularity Newman (2006a), profile-likelihood Bickel and Chen (2009), pseudo-likelihood Amini et al. (2013), variational methods Daudin, Picard and Robin (2008); Latouche, Birmele and Ambroise (2012), spectral clustering Rohe, Chatterjee and Yu (2011); Fishkind et al. (2013); Jin (2015), belief propagation Decelle et al. (2011), and so on. The asymptotic properties of these methods have also been established Bickel and Chen (2009); Rohe, Chatterjee and Yu (2011); Celisse, Daudin and Pierre (2012); Bickel et al. (2013); Gao, Lu and Zhou (2015); Zhang and Zhou (2016). However, most of these works assume that the number of communities k is known a priori. In real network data, k is usually unknown and needs to be estimated. Therefore, it is of importance to investigate how to choose k (called model selection in this article). Some methods have been proposed in recent years, including recursive approaches Zhao, Levina and Zhu (2011), spectral methodsLe and Levina (2015), sequential test Bickel and Sarkar (2015); Lei (2016), and network cross-validation Chen and Lei (2017). The likelihood methods for model selection have also been proposed Daudin, Picard and Robin (2008); Latouche, Birmele and Ambroise (2012); Saldana, Yu and Feng (2017). Wang and Bickel Wang and Bickel (2016) established the limiting distribution of the log-likelihood ratio under model misspecification–underfitting and overfitting, and deduced a likelihood-based model selection method. Specifically, their penalty function is

$$\lambda \frac{k(k+1)}{2} n \log n, \tag{1}$$

where λ is a tuning parameter. The penalty function $\frac{k(k+1)}{2}\log n$ (called the "BIC") was used to select the optimal number of communities in Daudin, Picard and Robin (2008) and Saldana, Yu and Feng (2017). When we use the penalty function (1) and the BIC to estimate k in simulations and real data examples, we find that they tend to underestimate and overestimate the number of communities, respectively. We propose a "corrected Bayesian information criterio" (CBIC) which is in the midway of those two criteria. Specifically, our penalty function is

$$\lambda n \log k + \frac{k(k+1)}{2} \log n, \tag{2}$$

which is smaller than that in (1) used by Wang and Bickel (2016). In comparison with the penalty function in the BIC used in Saldana, Yu and Feng (2017), it contains one more term $\lambda n \log k$. It is worth noting that Wang and Bickel Wang and Bickel (2016) dealt with the marginal log-likelihood where z as latent variables are integrated out, while here we plug a single estimated community assignment into the log-likelihood. These two log-likelihoods are equivalent asymptotically since the probability of the observations under incorrect community assignments is negligible asymptotically.

Similar to the classical BIC in linear and generalized linear models, the tuning parameter λ can be set to a constant, i.e., $\lambda = 1$. For the sake of convenience in the proofs and comparisons with Wang and Bickel (2016), we keep the tuning parameter λ in (2). Built on the work of Wang and Bickel (2016), we show that the proposed estimator is consistent. It is worth noting that we assume that the true number of communities is fixed in this article. Along the way of proving consistency, we obtain the Wilks theorem for stochastic block models. Furthermore, we show that, to obtain the consistency of model selection for stochastic block models, we need a so-called "consistency condition". We also provide sufficient conditions for both homogenous networks and non-homogenous networks. The results are further extended to degree corrected block models. The proposed method can be easily conducted by the popular spectral clustering methods Rohe, Chatterjee and Yu (2011); Jin (2015).

For the remainder of the paper, we proceed as follows. In Section 2, we re-examine the Bayesian paradigm for stochastic block models and propose the CBIC to determine the number of communities. In Section 3, we analyze the asymptotic behavior of the log-likelihood ratio and establish their asymptotic distributions. In Section 4, we establish the consistency of the estimator for the number of communities. We extend our results to degree corrected stochastic block models in Section 5. The numerical studies are given in Section 6. Some further discussions are made in Section 7. All proofs are given in Section 8.

2 Corrected BIC

In this section, we re-examine the Bayesian paradigm for the SBM and propose a corrected family of Bayesian information criteria.

For any fixed (θ, z), the log-likelihood of observing the adjacency matrix A under the stochastic block model up to a constant is

$$\log f(A|\theta, z) = \sum_{1\le a\le b\le k} (m_{ab} \log \theta_{ab} + (n_{ab} - m_{ab}) \log(1 - \theta_{ab})),$$

with count statistics

$$n_a = \sum_{i=1}^{n} \mathbf{1}\{z_i = a\}, \ n_{ab} = \sum_{i=1}^{n} \sum_{j\neq i} \mathbf{1}\{z_i = a, z_j = b\},$$

$$m_{ab} = \sum_{i=1}^{n} \sum_{j \neq i} A_{ij} \mathbf{1}\{z_i = a, z_j = b\}.$$

Daudin, Picard and Robin Daudin, Picard and Robin (2008) and Saldana, Yu and Feng Saldana, Yu and Feng (2017) used the following penalized likelihood function called the "BIC", to select the optimal number of communities:

$$\bar{\ell}(k) = \sup_{\theta \in \Theta_k} \log f(A|\theta, z) - \frac{k(k+1)}{2} \log n, \tag{3}$$

where $\Theta_k = [0,1]^{\frac{k(k+1)}{2}}$. According to our simulation studies, the BIC tends to overestimate the number of communities (see section 6). We now give some hint of why this phenomenon occurs. The BIC assumes that community assignment z is fixed for a given k. As soon as z is unknown and is estimated by some community detection methods, it cannot be treated as fixed any more. Therefore, the BIC essentially estimates the number of communities k using the penalized likelihood of the following form,

$$\tilde{\ell}(k) = \max_{z \in [k]^n} \sup_{\theta \in \Theta_k} \log f(A|\theta, z) - \frac{k(k+1)}{2} \log n.$$

Let Z be the set of all possible community assignments under consideration and let $\xi(z)$ be the prior probability of community assignment z. Assume that the prior density of θ is given by $\pi(\theta)$. Then the posterior probability of z is

$$P(z|A) = \frac{g(A|z)\xi(z)}{\sum_{z \in Z} g(A|z)\xi(z)},$$

where $g(A|z)$ is the likelihood of community assignment z, given by

$$g(A|z) = \int f(A|\theta, z)\pi(\theta)d\theta.$$

Under the Bayesian paradigm, a number of communities $\hat{k}$ that maximizes the posterior probability is selected. Since $\sum_{z \in Z} g(A|z)\xi(z)$ is a constant,

$$\hat{k} = \arg\max_{k} \max_{z \in [k]^n} g(A|z)\xi(z).$$

According to Daudin, Picard and Robin (2008), we have

$$\log g(A|z) = \sup_{\theta \in \Theta_k} \log f(A|\theta, z) - \frac{k(k+1)}{2} \log n + O(1).$$

Thus,

$$\log g(A|z)\xi(z) = \sup_{\theta \in \Theta_k} \log f(A|\theta, z) - \frac{k(k+1)}{2} \log n + O(1) + \log \xi(z). \tag{4}$$

By comparing equations (3) and (4), we can see that the BIC essentially assumes that $\xi(z)$

is constant for z over Z, i.e., $\xi(z) = 1/\tau(Z)$, where $\tau(Z)$ is the size of Z. Suppose that the number of nodes in the network is $n = 500$. The set of community assignments for $k = 2$, Z_2, has size 2^{500}, while the set of community assignments for $k = 3$, Z_3, has size 3^{500}. The constant prior in the BIC assigns probabilities to Z_k proportional to their sizes. Thus the probability assigned to Z_3 is 1.5^{500} times that assigned to Z_2. Community assignments with a larger number of communities get much higher probabilities than community assignments with fewer communities. This gives an explanation why the BIC tends to overestimate the number of communities.

This re-examination of the BIC naturally leads us to consider a new prior over Z. Assume that Z is partitioned into $\bigcup_{k=1} Z_k$. Let $\tau(Z_k)$ be the size of Z_k. We assign the prior distribution over Z in the following manner. We assign an equal probability to z in the same Z_k, i.e., $P(z|Z_k) = 1/\tau(Z_k)$ for any $z \in Z_k$. This is due to that all the community assignments in Z_k are equally plausible. Next, instead of assigning probabilities $P(Z_k)$ proportional to $\tau(Z_k)$, we assign $P(Z_k)$ proportional to $\tau^{-\delta}(Z_k)$ for some δ. Here $\delta > 0$ implies that small communities are plausible while $\delta < 0$ implies that large communities are plausible. This results in the prior probability

$$\xi(z) = P(z|Z_k)P(Z_k) \propto \tau^{-\lambda}(Z_k), \quad z \in Z_k,$$

where $\lambda = 1 + \delta$. Thus,

$$\log g(A|z)\xi(z) = \sup_{\theta \in \Theta_k} \log f(A|\theta, z) - \frac{k(k+1)}{2} \log n + O(1) - \lambda n \log k.$$

This type of prior distribution on the community assignment gives rise to a corrected BIC criterion as follows:

$$\ell(k) = \max_{z \in [k]^n} \sup_{\theta \in \Theta_k} \log f(A|\theta, z) - \left[\lambda n \log k + \frac{k(k+1)}{2} \log n \right], \tag{5}$$

where the second term is the penalty and $\lambda \geq 0$ is a tuning parameter. Then we estimate k by maximizing the penalized likelihood function:

$$\hat{k} = \arg\max_k \ell(k).$$

We make some remarks on the choice of the tuning parameter. If we have no prior information on the number of communities, i.e., both small communities and large communities are equally plausible, then $\lambda = 1$ ($\delta = 0$) is a good choice. It is similar to the case of variable selection in regression, where the BIC is also tuning free. We set $\lambda = 1$ in the CBIC for simulation studies which gives good performance.

In order to obtain the consistency of the CBIC, we analyze the asymptotic order of the log-likelihood ratio under model-misspecification in the next section.

3 Asymptotics of the log-likelihood ratio

In this section, we present the refined order of the log-likelihood ratio built on the work of Wang and Bickel (2016). The results here will be used for the proof of Theorem 4 in the next section.

We consider the following log-likelihood ratio

$$L_{k,k'} = \max_{z\in[k']^n} \sup_{\theta\in\Theta_{k'}} \log f(A|\theta, z) - \log f(A|\theta^*, z^*),$$

where θ^* and z^* are the true parameters. Further, k' is the number of communities under the alternative model and k is the true number of communities. Therefore, the comparison is made between the correct k-block model and a fitted k'-block model.

The asymptotic distributions of $L_{k,k'}$ for the cases $k' < k$ and $k' > k$ are given in this section. In Theorems 2.4 and 2.7 of Wang and Bickel (2016), the order of the log-likelihood ratios is $n^{3/2}$. Here, the order is n. We use the techniques developed in Wang and Bickel (2016) for the proofs of Theorems 1 and 3 with some refinements. For the case $k' = k$, we establish the Wilks theorem.

3.1 $k' < k$

We start with $k' = k - 1$. As discussed in Wang and Bickel (2016), a $(k-1)$-block model can be obtained by merging blocks in a k-block model. Specifically, given the true labels $z^* \in [k]^n$ and the corresponding block proportions $p = (p_1, \ldots, p_k)$, $p_a = n_a(z^*)/n$, we define a merging operation $U_{a,b}(\theta^*, p)$ which combines blocks a and b in θ^* by taking weighted averages with proportions in p. For example, for $\theta = U_{k-1,k}(\theta^*, p)$,

$$\theta_{ab} = \theta^*_{ab} \text{ for } 1 \le a, b \le k-2,$$

$$\theta_{a(k-1)} = \frac{p_a p_{k-1}\theta^*_{a(k-1)} + p_a p_k \theta^*_{ak}}{p_a p_{k-1} + p_a p_k} \text{ for } 1 \le a \le k-2,$$

$$\theta_{(k-1)(k-1)} = \frac{p^2_{k-1}\theta^*_{(k-1)(k-1)} + 2p_{k-1}p_k\theta^*_{(k-1)k} + p^2_k\theta^*_{kk}}{p^2_{k-1} + 2p_{k-1}p_k + p^2_k}.$$

For consistency, when merging two blocks (a, b) with $b > a$, the new merged block will be relabeled as a and all blocks c with $c > b$ will be relabeled as $c - 1$. Using this scheme, we also obtain the merged node labels $U_{a,b}(z)$ and the merged proportions $U_{a,b}(\theta^*, p)$.

Define

$$\mathcal{I} = \{(a, b) \in [k]^2 : k-1 \le a \le k \text{ or } k-1 \le b \le k\},$$

$$u(a) = \begin{cases} a, & \text{for } a \le k-2; \\ k-1, & \text{for } k-1 \le a \le k. \end{cases}$$

To obtain the asymptotic distribution of $L_{k,k'}$, we need the following condition.

(A1) There exists $C_1 > 0$ such that $\min_{1\le a\le k} n_a \ge C_1 n$ for all n.

The asymptotic distribution of $L_{k,k-1}$ is stated below, the proof of which is given in the Appendix.

Theorem 1. *Suppose that $A \sim P_{\theta^*,z^*}$ and condition (A1) holds. Then we have*

$$n^{-1}L_{k,k-1} - n\mu \xrightarrow{d} N(0, \sigma(\theta^*)),$$

where

$$\mu = \frac{1}{2n^2}\sum_{(a,b)\in\mathcal{I}} n_{ab}(\theta^*_{ab}\log\frac{\theta'_{u(a)u(b)}}{\theta^*_{ab}} + (1-\theta^*_{ab})\log\frac{1-\theta'_{u(a)u(b)}}{1-\theta^*_{ab}}),$$

$$\sigma(\theta^*) = \frac{1}{4}\sum_{(a,b)\in\mathcal{I}} C_{ab}\theta^*_{ab}(1-\theta^*_{ab})(\log\frac{\theta'_{u(a)u(b)}(1-\theta^*_{ab})}{(1-\theta'_{u(a)u(b)})\theta^*_{ab}})^2,$$

$$C_{ab} = \lim\frac{n_{ab}}{n^2}.$$

Condition (A1) requires that the size of each community is at least proportional to n. For finite k, it is satisfied almost surely if the membership vector σ is generated from a multinominal distribution with n trials and probability $\pi = (\pi_1, \ldots, \pi_k)$ such that $\min_{1\le i\le k}\pi_i \ge C_2$. For a general $k^- < k$, the same type of limiting distribution under condition (A1) still holds by assuming the uniqueness of the optimal merging scheme and identifiability after merging. But the proof will involve more tedious descriptions of how various merges can occur as discussed in Wang and Bickel (2016).

3.2 $k' = k$

In this case, we establish the Wilks theorem.

Theorem 2. *Suppose that $A \sim P_{\theta^*,z^*}$ and condition (A1) holds. Then we have*

$$2(\max_{z\in[k]^n}\sup_{\theta\in\Theta_k}\log f(A|\theta,z) - \log f(A|\theta^*,z^*)) \xrightarrow{d} \chi^2_{\frac{k(k+1)}{2}}.$$

3.3 $k' > k$

As discussed in Wang and Bickel (2016), it is very difficult to obtain the asymptotic distribution of $L_{k,k'}$ in the case of $k' > k$. Instead, we obtain its asymptotic order.

Theorem 3. *Suppose that $A \sim P_{\theta^*,z^*}$ and condition (A1) holds. Then we have*

$$\begin{aligned} L_{k,k^+} &\le \alpha n\log k^+ + \sup_{\theta\in\Theta_k}\log f(A|\theta,z^*) - \log f(A|\theta^*,z^*) \\ &= \alpha n\log k^+ + O_p(\tfrac{k(k+1)}{2}\log n), \end{aligned}$$

where $0 < \alpha \le 1 - \frac{C}{\log k^+} + \frac{2\log n + \log k}{n\log k^+}$.

4 Consistency of CBIC

In this section, we establish the consistency of the CBIC in the sense that it chooses the correct k with probability tending to one when n goes to infinity.

To obtain the consistency of the CBIC, we need an additional condition.

(A2) (Consistency condition) $n\mu \to -\infty$.

Theorem 4. *Suppose that $A \sim P_{\theta^*,z^*}$ and (A1)-(A2) hold. Let $\ell(k)$ be the penalized likelihood function for the CBIC, defined at (5).*
For $k' < k$,

$$P(\ell(k') > \ell(k)) \to 0.$$

For $k' > k$, when $\lambda > (\alpha \log k')/(\log k' - \log k)$,

$$P(\ell(k') > \ell(k)) \to 0,$$

where α is given in Theorem 3.

By Theorem 4, the probability $P(\ell(k') > \ell(k))$ goes to zero, regardless of the tuning parameter λ in the case of $k' < k$. When $k' > k$, it depends on the parameter λ. Then a natural question is whether the choice of $\lambda = 1$ is good. Note that it also depends on α. With an appropriate α, the probability $P(\ell(k') > \ell(k))$ also goes to zero when $\lambda = 1$ as demonstrated in the following corollary.

Corollary 5. *Suppose that $A \sim P_{\theta^*,z^*}$ and (A1)-(A2) hold. Let $\ell(k)$ be the penalized likelihood function for the CBIC, defined at (5).*
For $k' < k$,

$$P(\ell(k') > \ell(k)) \to 0.$$

For $k' > k$, suppose $\alpha < 1 - \frac{\log k}{\log k'}$, for $\lambda = 1$,

$$P(\ell(k') > \ell(k)) \to 0.$$

By checking the proof of Theorem 4, it is easy to see that for $k' > k$, $P(\tilde{\ell}(k') > \tilde{\ell}(k)) \to 1$. This implies that the BIC tends to overestimate the number of communities k for stochastic block models.

Remark 1. Note that in the proof of Theorem 4, for $k' < k$, every step is reversible. Thus, for appropriate λ, under the condition $(A1)$, the consistency condition $n\mu \to -\infty$ is a necessary and sufficient condition.

Next we analyze that under what condition we have $n\mu \to -\infty$. We consider both homogeneous network and non-homogeneous network. Consider a homogeneous network $\theta^* = (\theta^*_{ab})_{1\le a\le b\le k}$, where $\theta^*_{aa} = p$ for all $a = 1, \dots, k$ and $\theta^*_{ab} = q$ for $1 \le a < b \le k$. For simplicity, we assume $n_{ab}/n^2 = C_3$. That is,

$$\mu = \frac{C_3}{2} \sum_{(a,b)\in\mathcal{I}} (\theta^*_{ab} \log \frac{\theta'_{u(a)u(b)}}{\theta^*_{ab}} + (1 - \theta^*_{ab}) \log \frac{1 - \theta'_{u(a)u(b)}}{1 - \theta^*_{ab}}).$$

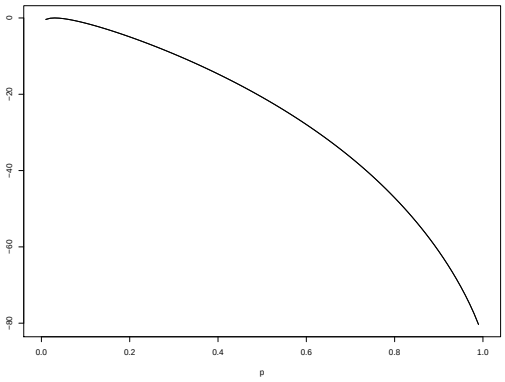

Figure 1: $n\mu$ vs p

Set $q = 0.03$, $C_3 = 0.2$, $n = 500$, $x_1 = \frac{n_{(k-1)(k-1)}}{n_{(k-1)(k-1)}+n_{(k-1)k}} = 0.5$. We can see from Figure 1 that under the condition that $p/q \geq 6$, we have $n\mu \to -\infty$. We also changed the values of x_1 and C_3 and found that the result is similar.

In fact, for a homogeneous network, we have the following result.

Theorem 6. *Suppose that p/q is sufficiently large, then $n\mu \to -\infty$.*

Here p/q is sufficiently large means that there exists a constant $C > 1$ such that $p/q \geq C$.

For a non-homogeneous network, we have the similar result:

Corollary 7. *Suppose that $min_{1\leq a<b\leq k}(\theta^*_{aa}/\theta^*_{ab})$ is sufficiently large, then $n\mu \to -\infty$.*

Remark 2. We notice that the tuning parameter λ being equal to 1 may not always be the best choice. Simulation studies in Section 6 shed some lights. In our simulations, we found that $\lambda = 1$ have a good performance for estimating the number of communities k in the following two cases: (1) when the number of communities k is small (e.g., $k \leq 4$), for both large p/q (e.g., $p/q \geq 6$) and medium p/q (e.g., $p/q \geq 4$); (2) when the number of communities k is large (e.g., $k \geq 5$), for large p/q (e.g., $p/q \geq 6$). However, when the number of communities k is large (e.g., $k \geq 5$), for both medium and small p/q (e.g., $1 < p/q \leq 4$), $\lambda = 1$ tends to underestimate the number of communities k. As a result, $0 \leq \lambda < 1$ may be a better choice. For this case, we may use the cross-validation method to choose the tuning parameter λ. We would like to explore this problem in future work.

5 Extension to a degree-corrected SBM

Real-world networks often include a number of high degree "hub" nodes that have many connections Barabási and Bonabau (2003). To incorporate the degree heterogeneity within communities, the degree corrected stochastic block model (DCSBM) was proposed by Karrer and Newman (2011). Specifically, this model assumes that $P(A_{ij} = 1 \mid z(i), z(j)) = \omega_i\omega_j\theta_{z(i)z(j)}$, where $\omega = (\omega_i)_{1\leq i\leq n}$ are a set of node degree parameters measuring the degree variation within

blocks. For identifiability of the model, we need the constraint $\sum_i \omega_i \mathbf{1}\{z(i) = a\} = n_a$ for each community $1 \le a \le k$.

Similar to Karrer and Newman (2011), we replace the Bernoulli random variables A_{ij} by the Poisson random variable. As discussed in Zhao, Levina and Zhu (2012), there is no practical difference in performance. The reason is that the Bernoulli distribution with a small mean is well approximated by the Poisson distribution. An advantage using Poisson distributions is that it will greatly simplify the calculations. Another advantage is that it will allow networks containing both multi-edges and self-edges.

For any fixed (θ, ω, z), the log-likelihood of observing the adjacency matrix A under the degree corrected stochastic block model is

$$\log f(A|\theta, \omega, z) = \sum_{1 \le i \le n} d_i \log \omega_i + \sum_{1 \le a \le b \le k} (m_{ab} \log \theta_{ab} - n_{ab}\theta_{ab}),$$

where $d_i = \sum_{1 \le j \le n} A_{ij}$.

We first consider the case ω is known, which was also assumed by Lei (2016), Gao, Ma, Zhang and Zhou (2016) in their theoretical analyses. With similar arguments, one can show that the previous Theorems 1 and 3 still hold in the DCSBM. Although Theorem 2 does not hold in the DCSBM, we have the following result.

Theorem 8. *Suppose that $A \sim \mathbb{P}_{\theta^*, z^*}$ and (A1) holds. Then we have*

$$\max_{z \in [k]^n} \sup_{\theta \in \Theta_k} \log f(A|\theta, \omega, z) - \log f(A|\theta^*, \omega, z^*) = O_p(\frac{k(k+1)}{2} \log n).$$

Therefore, Theorem 4 still holds in the DCSBM.

If ω_i's are unknown, we use the plug-in method. That is, we need to estimate ω_i's. After allowing for the identifiability, constrain on ω, the MLE of the parameter ω_i is given by $\hat{\omega}_i = n_a d_i / \sum_{j: z_j = z_i} d_j$. Simulations show that the CBIC can estimate k with high accuracy for the DCSBM.

6 Experiments

6.1 Algorithm

Since there are k^n possible assignments for the communities, it is intractable to directly optimize the log-likelihood of the SBM. Since the primary goal of our article is to study the penalty function, we use a computationally feasible algorithm–spectral clustering to estimate the community labels for a given k.

The algorithm finds the eigenvectors $u_1, \ldots, u_k$ associated with the k eigenvalues of the Laplacian matrix that are largest in magnitude, forming an $n \times k$ matrix $U = (u_1, \ldots, u_k)$, and then applies the k-means algorithm to the rows of U. For details, see Rohe, Chatterjee and Yu (2011). They established the consistency of spectral clustering in the stochastic block model

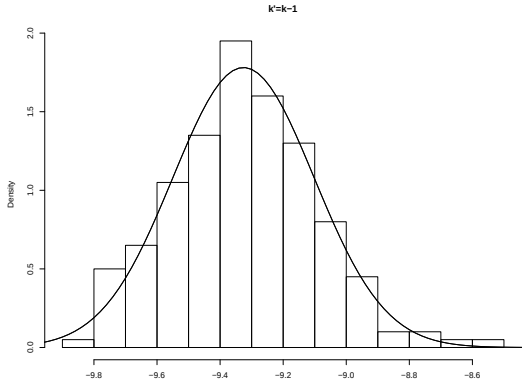

Figure 2: Emipirical distribution of $n^{-1}L_{k,k-1}$. The solid curve is normal density with mean $n\mu$ and $\sigma(\theta^*)$ as given in Theorem 1.

under proper conditions imposed on the density of the network and the eigen-structure of the Laplacian matrix.

For the DCSBM, we apply a variant of spectral clustering, called spectral clustering on ratios-of-eigenvectors (SCORE) proposed by Jin (2015). Instead of using the Laplacian matrix, the SCORE collects the eigenvectors $v_1, \ldots, v_k$ associated with the k eigenvalues of A that are largest in magnitude, and then forms the $n \times k$ matrix $V = (\mathbf{1}, v_2/v_1, \ldots, v_k/v_1)$. The SCORE then applies the k-means algorithm to the rows of V. The corresponding consistency results for the DCSBM are also provided.

We restrict our attention to candidate values for the true number of communities in the range $k' \in \{1, \ldots, 18\}$, both in simulations and the real data analysis.

6.2 Simulations

Simulation 1. In the SBM setting, we first compare the empirical distribution of the log-likelihood ratio with the asymptotic results on Theorems 1, 2 and 3. We set the network size as $n = 500$ and the probability matrix $\theta^*_{ab} = 0.03(1 + 5 \times \mathbf{1}(a = b))$. We set $n = 500$ and $k = 3$ with $\pi_1 = \pi_2 = \pi_3 = 1/3$. Each simulation in this section is repeated 200 times. The plot for $n^{-1}L_{k,k-1}$ is shown in Figure 2. The empirical distribution is well approximated by the normal distribution in the case of underfitting. Figure 3 plots the empirical distribution of $2L_{k,k}$ in the case of $k' = k$. The distribution also matches the chi-square distribution well. Figure 4 plots the empirical distribution of $L_{k,k+1}$.

Simulation 2. In the SBM setting, we investigate how the accuracy of the CBIC changes as the tuning parameter λ varies. We let λ increase from 0 to 3.5. The probability matrix is the same as in Simulation 1. We set each block size according to the sequence $(60, 90, 120, 150, 60, 90, 120, 150)$. That is, if $k = 1$, we set the network size n to be 60; if $k = 2$, we set two respective block sizes to be 60 and 90; and so forth. This setting is the same as in Saldana, Yu and Feng (2017). As can be seen in Figure 5, the rate of the successful recovery of the number of communities is very low when λ is close to zero. When λ is between 0.5 and 1.5, the success rate is almost with 100%;

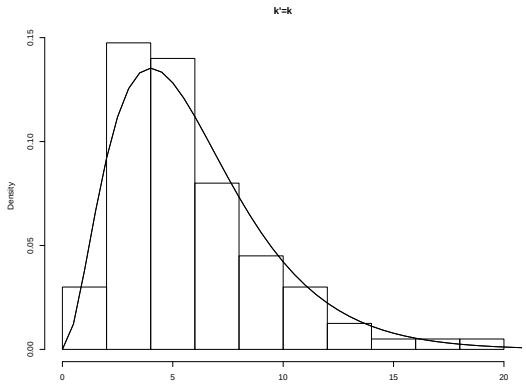

Figure 3: Emipirical distribution of $2L_{k,k}$. The solid curve is chi-square density with degree $\frac{k(k+1)}{2} = 6$.

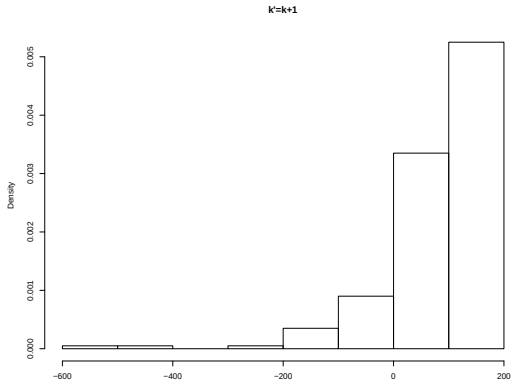

Figure 4: Emipirical distribution of $2L_{k,k+1}$.

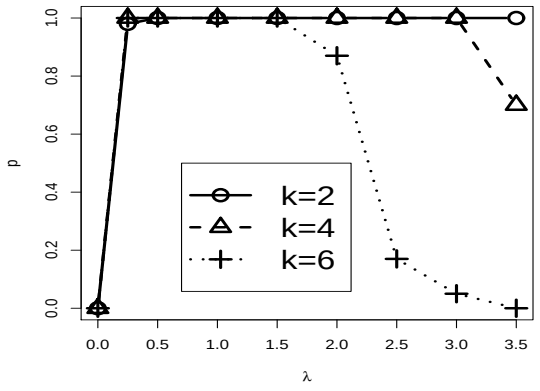

Figure 5: Success rate vs λ.

When λ becomes larger, the success rate decreases in which the change point depends on k. It can be seen from Figure 5 that $\lambda = 1$ is a good tuning parameter.

Simulation 3. In the SBM setting, we compare the CBIC with the BIC proposed by Daudin, Picard and Robin (2008) and Saldana, Yu and Feng (2017) and the bootstrap corrected sequential test proposed by Lei (2016). For the bootstrap corrected sequential test, we select threshold t_n corresponding to nominal type I error bound 10^{-4}. The network size n is the same as in Simulation 2 and the probability matrix is $\theta^*_{ab} = 0.03(1 + r \times \mathbf{1}(a = b))$. The numerical results are shown in Tables 1, 2 and 3. From these Tables, we can see that the CBIC shows a significant improvement over the BIC and the bootstrap corrected sequential test. It can be seen from Table 1 that, for large p/q (e.g., $r = 5$), the CBIC recovers the number of communities k perfectly while the success rates for the BIC and the bootstrap corrected sequential test are low for $k \leq 4$ and $k \geq 5$, respectively. It can also be seen from Table 3 that, for medium p/q (e.g., $r = 3$), the CBIC recovers the number of communities k quite well for $k \leq 4$. When the number of communities k is large (e.g., $k \geq 5$), for medium p/q (e.g., $r = 3$), the BIC outperforms the CBIC. For this case, the performance of the CBIC may be improved by tuning λ using cross-validation. We also implemented the penalized likelihood method (PLH) proposed by Wang and Bickel (2016) and found out that it usually underestimates the number of communities k.

Table 1: performance of CBIC for SBM: r=5

	CBIC ($\lambda = 1$)		BIC		Lei (2016)	
	Prob	Mean (Var)	Prob	Mean (Var)	Prob	Mean (Var)
$k = 2$	1	2 (0)	0.19	3.11 (0.70)	1	2 (0)
$k = 3$	1	3 (0)	0.48	3.64 (0.56)	0.99	3.03 (0.09)
$k = 4$	1	4 (0)	0.72	4.36 (0.45)	0.97	4.02 (0.02)
$k = 5$	1	5 (0)	0.92	5.09 (0.10)	0.86	4.86 (0.12)
$k = 6$	1	6 (0)	1	6.00 (0.00)	0.65	5.65 (0.23)
$k = 7$	1	7 (0)	1	7.00 (0.00)	0.21	6.21 (0.17)
$k = 8$	1	8 (0)	1	8.00 (0.00)	0.16	7.16 (0.14)

Table 2: performance of CBIC for SBM: r=4

	CBIC ($\lambda = 1$)		BIC		Lei (2016)	
	Prob	Mean (Var)	Prob	Mean (Var)	Prob	Mean (Var)
$k = 2$	1	2 (0)	0.32	3.14 (1.23)	1	2 (0)
$k = 3$	1	3 (0)	0.70	3.42 (0.57)	0.95	3.04 (0.04)
$k = 4$	1	4 (0)	0.85	4.16 (0.16)	0.62	4.31 (0.97)
$k = 5$	0.97	4.97 (0.02)	0.92	5.09 (0.10)	0.08	4.50 (0.76)
$k = 6$	0.93	5.93 (0.04)	1	6.00 (0.00)	0.02	5.14 (0.31)
$k = 7$	0.90	6.9 (0.09)	1	7.00 (0.00)	0.04	6.00 (0.49)
$k = 8$	0.84	7.84 (0.13)	0.98	8.02 (0.02)	0.05	6.40 (0.65)

Simulation 4. In the SBM setting, we compare the CBIC with the BIC and the bootstrap corrected sequential test for a non-homogeneous network. The block sizes are $(60, 90, 120, 150)$

Table 3: performance of CBIC for SBM: r=3

	CBIC ($\lambda = 1$)		BIC		Lei (2016)	
	Prob	Mean (Var)	Prob	Mean (Var)	Prob	Mean (Var)
$k = 2$	1	2 (0)	0.19	3.11 (0.70)	0.99	2.01 (0.01)
$k = 3$	0.98	2.98 (0.19)	0.31	3.36 (1.60)	0.47	3.00 (1.42)
$k = 4$	0.91	3.91 (0.08)	0.92	4.06 (0.16)	0.15	3.33 (0.81)
$k = 5$	0.33	4.22 (0.39)	0.98	5.02 (0.02)	0.13	3.67 (1.43)
$k = 6$	0.16	4.94 (0.40)	0.87	5.91 (0.12)	0.07	4.17 (0.28)
$k = 7$	0.12	5.77 (0.46)	0.68	6.68 (0.23)	0.00	4.89 (0.20)
$k = 8$	0.03	6.34 (0.41)	0.41	7.48 (0.39)	0.00	5.50 (0.25)

and the probability matrix is

$$\theta^* = \rho \begin{pmatrix} 0.20 & 0.04 & 0.05 & 0.03 \\ 0.03 & 0.20 & 0.03 & 0.05 \\ 0.05 & 0.03 & 0.25 & 0.04 \\ 0.03 & 0.05 & 0.04 & 0.25 \end{pmatrix}.$$

Here $k = 4$ is small and $\min_{1\le a<b\le k}(\theta^*_{aa}/\theta^*_{ab}) = 4$ is medium. The numerical results are given in Table 4. The comparisons are similar to those in Tables 1, 2 and 3.

Table 4: performance of CBIC for SBM: non-homogeneous

	CBIC ($\lambda = 1$)		BIC		Lei (2016)	
	Prob	Mean (Var)	Prob	Mean (Var)	Prob	Mean (Var)
$\rho = 0.7$	0.98	3.99 (0.01)	0.83	4.17 (0.21)	0.12	4.52 (5.15)
$\rho = 0.8$	1	4 (0)	0.82	4.21 (0.24)	0.25	5.40 (1.67)
$\rho = 0.9$	1	4 (0)	0.77	4.23 (0.18)	0.50	5.50 (2.90)
$\rho = 1.0$	1	4 (0)	0.80	4.32 (0.15)	0.86	4.28 (0.57)
$\rho = 1.2$	1	4 (0)	0.72	4.34 (0.17)	0.95	4.09 (0.19)

Simulation 5. In the DCSBM setting, we investigate the performance of the CBIC for the DCSBM. Since the bootstrap corrected sequential test is only designed for the SBM, we compare the CBIC with the BIC and the network cross-validation proposed by Chen and Lei (2017). In choosing the parameters θ, ω in the DCSBM, we follow the approach proposed in Zhao, Levina and Zhu (2012). ω_i is independently generated from a distribution with expectation 1, that is

$$\omega_i = \begin{cases} \eta_i, & \text{w.p. } 0.8; \\ 7/11, & \text{w.p. } 0.1; \\ 15/11, & \text{w.p. } 0.1, \end{cases}$$

where η_i is uniformly distributed on the interval $[\frac{3}{5}, \frac{7}{5}]$. The edge probability and network sizes are set the same as in Simulation 3. The numerical results are given in Tables 5, 6 and 7. The

comparisons are similar to those in Tables 1, 2 and 3.

Table 5: performance of CBIC for DCSBM: r=5

	CBIC ($\lambda = 1$)		BIC		NCV	
	Prob	Mean (Var)	Prob	Mean (Var)	Prob	Mean (Var)
$k = 2$	0.97	2.04 (0.06)	0	4.35 (1.62)	0.94	2.14 (0.40)
$k = 3$	1	3 (0)	0.11	5.38 (2.38)	0.94	3.06 (0.06)
$k = 4$	1	4 (0)	0.29	5.67 (2.34)	0.89	4.14 (0.18)
$k = 5$	0.97	5.06 (0.09)	0.20	7.07 (2.71)	0.33	5.09 (0.66)
$k = 6$	0.98	6.02 (0.02)	0.52	7.00 (1.61)	0.29	7.41 (1.58)
$k = 7$	0.93	7.06 (0.09)	0.57	7.63 (0.76)	0.25	8.50 (1.26)
$k = 8$	0.95	8.05 (0.05)	0.58	8.59 (0.61)	0.15	9.38 (0.54)

Table 6: performance of CBIC for DCSBM: r=4

	CBIC ($\lambda = 1$)		BIC		NCV	
	Prob	Mean (Var)	Prob	Mean (Var)	Prob	Mean (Var)
$k = 2$	0.96	2.04 (0.39)	0.03	4.57 (1.78)	0.80	2.52 (1.56)
$k = 3$	0.97	3.03 (0.29)	0.11	5.18 (1.93)	0.65	3.60 (1.25)
$k = 4$	0.93	4.09 (0.12)	0.42	5.27 (2.26)	0.14	4.12 (3.52)
$k = 5$	0.72	5.05 (0.43)	0.35	6.32 (1.81)	0.17	5.42 (3.71)
$k = 6$	0.50	5.70 (0.62)	0.56	6.84 (1.53)	0.16	6.41 (2.27)
$k = 7$	0.43	6.76 (0.73)	0.44	7.85 (1.79)	0.23	8.00 (2.02)
$k = 8$	0.24	7.39 (0.56)	0.37	8.61 (0.88)	0	8.76 (1.43)

6.3 Real data analysis

6.3.1 International trade dataset

We study an international trade dataset collected by Westveld and Hoff (2011). It contains yearly international trade data among $n = 58$ countries from 1981–2000. One can refer to Westveld and Hoff (2011) for a detail description. This dataset was revisited by Saldana, Yu and Feng (2017) for the purpose of estimating the number of communities. Following their paper, we only focus on data from 1995 and transform the weighted adjacency matrix to the binary matrix using their methods. An adjacency matrix A is created by first considering a weight matrix W with $W_{ij} = \text{Trade}_{ij} + \text{Trade}_{ji}$, where Trade_{ij} denotes the value of exports from country i to country j. Define $A_{ij} = 1$ if $W_{ij} \geq W_\alpha$, and $A_{ij} = 0$ otherwise. Here W_α denotes the α-th quantile of $\{W_{ij}\}_{1 \leq i < j \leq n}$. We set $\alpha = 0.5$ as in Saldana, Yu and Feng (2017). At $\lambda = 1$, the CBIC for the SBM estimates $\hat{k} = 5$, while the BIC and the NCV estimate $\hat{k} = 10$ and $\hat{k} = 3$, respectively. The CBIC for the DCSBM estimates $\hat{k} = 3$, while both the BIC and the NCV estimate $\hat{k} = 1$. As discussed in Saldana, Yu and Feng (2017), it is seems reasonable to select 3 communities, corresponding to countries with highest GDPs, industrialized European

Table 7: performance of CBIC for DCSBM: r=3

	CBIC ($\lambda = 1$)		BIC		NCV	
	Prob	Mean (Var)	Prob	Mean (Var)	Prob	Mean (Var)
$k = 2$	0.97	2.02 (0.03)	0	5.02 (1.80)	0.74	2.44 (1.50)
$k = 3$	1	3 (0)	0.19	5.09 (3.03)	0.16	4.20 (6.73)
$k = 4$	0.33	3.55 (0.53)	0.15	6.16 (3.18)	0.15	3.56 (1.80)
$k = 5$	0.10	4.12 (0.67)	0.14	6.23 (3.91)	0.13	3.82 (3.74)
$k = 6$	0.09	4.86 (0.61)	0.21	6.44 (2.65)	0.10	5.56 (5.98)
$k = 7$	0.06	5.54 (0.60)	0.18	6.61 (1.93)	0.07	6.43 (5.88)
$k = 8$	0	6.09 (0.12)	0.13	7.07 (1.34)	0	9.09 (3.02)

and Asian countries with medium-level GDPs, and developing countries in South America with the smallest GDPs.

6.3.2 Political blog dataset

We study the political blog network Adamic and Glance (2005), collected around 2004. This network consists of blogs about US politics, with edges representing web links. The nodes are labeled as "conservative" and "liberal" by the authors of Adamic and Glance (2005). So it is reasonable to assume that this networks contains these two communities. We only consider its largest connected component of this network which consists of 1222 nodes with community sizes 586 and 636 as is commonly done in the literature. It is widely believed that the DCSBM is a better fit for this network than the SBM. At $\lambda = 1$, the CBIC for the DCSBM estimates $\hat{k} = 2$, while the PLH and the NCV estimate $\hat{k} = 1$ and $\hat{k} = 2$, respectively. We can see that both the CBIC and the NCV give a reasonable estimate for the number of communities.

7 Discussion

In this paper, under both the SBM and the DCSBM, we have proposed a "corrected Bayesian information criterion" that leads a consistent estimator for the number of communities. The criterion improves those used in Wang and Bickel (2016) and Saldana, Yu and Feng (2017) which tend to underestimate and overestimate the number of communities, respectively. The simulation results indicate that the criterion has a good performance for estimating the number of communities for finite sample sizes.

Some extensions may be possible. For instance, it is interesting to study whether the CBIC is still consistent for correlated binary data. For this case, it seems that we can consider the composite likelihood studied in Saldana, Yu and Feng (2017). In addition, we only prove the asymptotic results in the case that the number of communities is fixed. It is interesting to explore whether our results can be extended to high dimensional SBMs. We have also noticed that $\lambda = 1$ is not always the best choice. When the number of communities k is large (e.g., $k \geq 5$), for both medium and small p/q (e.g., $1 < p/q \leq 4$), $\lambda = 1$ tends to underestimate the

number of communities k. As a result, $0 \leq \lambda < 1$ may be a better choice. For this case, we may use the cross-validation method to choose the tuning parameter λ, which will be explored in future work.

8 Appendix

We quote some notations from Wang and Bickel (2016). Define

$$F(M,t) = \sum_{a,b=1}^{k'} t_{ab}\gamma(\frac{M_{ab}}{t_{ab}}),$$

where $\gamma(x) = x\log x + (1-x)\log(1-x)$. Then

$$\sup_{\theta\in\Theta_{k'}} \log f(A|\theta,z) = \frac{n^2}{2}F(\frac{m(z)}{n^2},\frac{n(z)}{n^2}).$$

Define

$$G(R(z),\theta^*) = \sum_{a,b=1}^{k'} [R\mathbf{1}\mathbf{1}^T R^T(z)]_{ab}\gamma(\frac{R\theta^* R^T(z)}{R\mathbf{1}\mathbf{1}^T R^T(z)}),$$

where $R(z')$ is the $k'\times k$ confusion matrix whose (a,b)-entry is

$$R_{ab}(z',z^*) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}\{z'_i = a, z^*_j = b\}.$$

8.1 Proofs for Theorem 1

Without loss of generality assume the maximum of $G(R(z),\theta^*)$ is achieved at $z' = U_{k-1,k}(z)$. Denote $\theta' = U_{k-1,k}(\theta^*,p^*)$. The following Lemmas 1 and 2 are from Wang and Bickel (2016), which will be used in the proof of Theorem 1.

Lemma 1. *Given the true labels z^* with block proportions $p = n(z^*)/n$, maximizing the funciton $G(R(z),\theta^*)$ over R achieves its maximum in the label set*

$$\{z\in[k-1]^n : \textit{there exists } \tau \textit{ such that } \tau(z) = U_{ab}(z), 1\leq a<b\leq k\},$$

where U_{ab} merges z with labels a and b.

Furthermore, suppose z' gives the unique maximum (up to a permutation τ), for all R such that $R\geq 0$, $R^T\mathbf{1} = p$,

$$\frac{\partial G((1-\epsilon)R(z') + \epsilon R(z),\theta^*)}{\partial\epsilon}\,|_{\epsilon=0^+} < -C_2 < 0.$$

Lemma 2. *Suppose* $z \in [k']^n$ *and define* $X(z) = \frac{m(z)}{n^2} - R\theta^* R(z)^T$. *For* $\epsilon \leq 3$,

$$P(\max_{z\in[k']^n} \parallel X(z) \parallel \geq \epsilon) \leq 2(k')^{n+2} \exp(-\frac{n^2\epsilon^2}{4(\parallel \theta^* \parallel_\infty +1)}).$$

Let $y \in [k']^n$ *be a fixed set of labels, then for* $\epsilon \leq \frac{3m}{n}$,

$$\begin{aligned} &P(\max_{z:|x-y|\leq m} \parallel X(z) - X(y) \parallel_\infty > \epsilon\tfrac{m}{n}) \\ &\leq 2\tbinom{n}{m}(k')^{m+2} \exp(-\tfrac{n^3\epsilon^2}{4m(4\|\theta^*\|_\infty+1)}). \end{aligned}$$

In order to prove Theorem 1, we need one lemma below.

Lemma 3. *Suppose that* $A \sim P_{\theta^*,z^*}$. *With probability tending to 1,*

$$\max_{z\in[k-1]^n} \sup_{\theta\in\Theta_{k-1}} \log f(A|\theta, z) = \sup_{\theta\in\Theta_{k-1}} \log f(A|\theta, z')$$

Proof. The arguments are similar to those for Lemma 2.3 in Wang and Bickel (2016). By Lemma 1, suppose $G(R(z), \theta^*)$ is maximized at z'. Define

$$I^-_{\delta_n} = \{z \in [k-1]^n : G(R(z), \theta^*) - G(R(z'), \theta^*) < -\delta_n\},$$

for $\delta_n \to 0$ slowly enough.

By Lemma 2, for $\epsilon_n \to 0$ slowly,

$$\mid F(m(z)/n^2, n(z)/n^2) - G(R(z), \theta^*) \mid \leq C \sum_{a,b=1}^{k-1} \mid m_{ab}(z)/n^2 - (R\theta^* R^T(z))_{ab} \mid = o_p(\epsilon_n)$$

since γ is Lipschitz on any interval bounded away from 0 and 1.

$$\begin{aligned} & \max_{z\in I^-_{\delta_n}} \sup_{\theta\in\Theta_{k-1}} \log f(A|\theta, z) \\ \leq\ & \log(\textstyle\sum_{z\in I^-_{\delta_n}} \sup_{\theta\in\Theta_{k-1}} f(A|\theta, z)) \\ =\ & \log(\textstyle\sum_{z\in I^-_{\delta_n}} \sup_{\theta\in\Theta_{k-1}} e^{\log f(A|\theta,z)}) \\ \leq\ & \log(\sup_{\theta\in\Theta_{k-1}} f(A|\theta, z')(k-1)^n e^{o_p(n^2\epsilon_n)-n^2\delta_n}) \\ =\ & \log(\sup_{\theta\in\Theta_{k-1}} f(A|\theta, z')) + \log((k-1)^n e^{o_p(n^2\epsilon_n)-n^2\delta_n}) \\ <\ & \sup_{\theta\in\Theta_{k-1}} \log f(A|\theta, z'). \end{aligned}$$

For $z \notin I^-_{\delta_n}$, $G(R(z), \theta^*) - G(R(z'), \theta^*) \to 0$. Let $\bar{z} = \min \mid \tau(z) - z' \mid$. Since the maximum is unique up to τ, $\parallel R(\bar{z}) - R(z') \parallel_\infty \to 0$.

By Lemma 2,

$$\begin{aligned} & P(\max_{z\notin\tau(z')} \parallel X(\bar{z}) - X(z') \parallel_\infty > \epsilon \mid z - z' \mid /n) \\ \leq\ & \textstyle\sum_{m=1}^n P(\max_{z:z=\bar{z},|\bar{z}-z'|=m} \parallel X(\bar{z}) - X(z') \parallel_\infty > \epsilon\frac{m}{n}) \\ \leq\ & \textstyle\sum_{m=1}^n 2(k-1)^{k-1} n^m (k-1)^{m+2} e^{-Cnm} \to 0. \end{aligned}$$

It follows for $\mid z - z' \mid = m$, $z \notin I^-_{\delta_n}$,

$$\begin{aligned} \| \tfrac{m(\bar{z})}{n^2} - \tfrac{m(z')}{n^2} \|_\infty &= o_p(1)\tfrac{|z-z'|}{n} + \| R\theta^* R^T(\bar{z}) - R\theta^* R^T(z') \|_\infty \\ &\geq \tfrac{m}{n}(C + o_p(1)). \end{aligned}$$

Observe that $\| \frac{m(z')}{n^2} - R\theta^* R^T(z') \|_\infty = o_p(1)$. By Lemma 2, $\| \frac{n(z')}{n^2} - R\mathbf{1}\mathbf{1}^T R^T(z') \|_\infty = o_p(1)$. Note that $F(\cdot,\cdot)$ has continuous derivation in the neighborhood of $(\frac{m(z')}{n^2}, \frac{n(z')}{n^2})$. By Lemma 1,

$$\frac{\partial F((1-\epsilon)\frac{m(z')}{n^2} + \epsilon M, (1-\epsilon)\frac{n(z')}{n^2} + \epsilon t)}{\partial \epsilon} \mid_{\epsilon = 0^+} < -C < 0$$

for (M, t) in the neighborhood of $(\frac{m(z')}{n^2}, \frac{n(z')}{n^2})$. Thus,

$$F(\frac{m(\bar{z})}{n^2}, \frac{n(\bar{z})}{n^2}) - F(\frac{m(z')}{n^2}, \frac{n(z')}{n^2}) \leq -C\frac{m}{n}.$$

Since

$$\begin{aligned} & \sup_{\theta \in \Theta_{k-1}} \log f(A|\theta, z) - \sup_{\theta \in \Theta_{k-1}} \log f(A|\theta, z') \\ \leq\ & n^2 (F(\tfrac{m(\bar{z})}{n^2}, \tfrac{n(\bar{z})}{n^2}) - F(\tfrac{m(z')}{n^2}, \tfrac{n(z')}{n^2})) \\ =\ & -Cmn, \end{aligned}$$

we have

$$\begin{aligned} & \max_{z \notin I^-_{\delta_n}, z \notin \tau(z')} \sup_{\theta \in \Theta_{k-1}} \log f(A|\theta, z) \\ \leq\ & \log(\textstyle\sum_{z \notin I^-_{\delta_n}, z \notin \tau(z')} \sup_{\theta \in \Theta_{k-1}} f(A|\theta, z)) \\ \leq\ & \log(\textstyle\sum_{z \in \tau(z')} \sup_{\theta \in \Theta_{k-1}} f(A|\theta, z) \sum_{m=1}^n (k-1)^m n^m e^{-Cmn}) \\ \leq\ & \log(\sup_{\theta \in \Theta_{k-1}} f(A|\theta, z') \textstyle\sum_{z \in \tau(z')} \sum_{m=1}^n (k-1)^m n^m e^{-Cmn}) \\ =\ & \sup_{\theta \in \Theta_{k-1}} \log f(A|\theta, z') + \log((k-1)^{k-1} \textstyle\sum_{m=1}^n (k-1)^m n^m e^{-Cmn}) \\ <\ & \sup_{\theta \in \Theta_{k-1}} \log f(A|\theta, z'). \end{aligned}$$

□

Proof of Theorem 1. By Hoeffding's (1963) inequality, we have

$$\begin{aligned} P(\max_{1 \leq a \leq b \leq k} \mid \theta^*_{ab} - \hat{\theta}_{ab} \mid > t) &\leq \textstyle\sum_{1 \leq a \leq b \leq k} P(\mid \theta^*_{ab} - \hat{\theta}_{ab} \mid > t) \\ &\leq k^2 e^{-2t^2 n_a n_b} \\ &\leq e^{2\log k - 2t^2 n_a n_b} = e^{2\log k - 2C_1^2 n^2 t^2}. \end{aligned}$$

It implies that

$$\max_{1 \leq a \leq b \leq k} \mid \theta^*_{ab} - \hat{\theta}_{ab} \mid = o_p(\frac{\log n}{n}).$$

Note that $\sup_{\theta \in \Theta_{k-1}} \log f(A|\theta, z')$ is uniquely maximized at

$$\hat{\theta}_{ab} = \frac{m_{ab}}{n_{ab}} = \theta^*_{ab} + o_p(\frac{\log n}{n}) \text{ for } 1 \leq a, b \leq k-2,$$

$$\hat{\theta}_{a(k-1)} = \frac{m_{a(k-1)} + m_{ak}}{n_{a(k-1)} + n_{ak}} = \theta'_{a(k-1)} + o_p(\frac{\log n}{n}) \text{ for } 1 \le a, b \le k-2,$$

$$\hat{\theta}_{(k-1)(k-1)} = \frac{\sum_{a=k-1}^{k}\sum_{b=a}^{k} m_{ab}}{\sum_{a=k-1}^{k}\sum_{b=a}^{k} n_{ab}} = \theta'_{(k-1)(k-1)} + o_p(\frac{\log n}{n}).$$

Therefore, we have

$$\begin{aligned}
& n^{-1}(\max_{z\in[k-1]^n}\sup_{\theta\in\Theta_{k-1}}\log f(A|\theta,z) - \log f(A|\theta^*,z^*)) \\
= \; & n^{-1}(\sup_{\theta\in\Theta_{k-1}}\log f(A|\theta,z') - \log f(A|\theta^*,z^*)) \\
= \; & n^{-1}(\tfrac{1}{2}\sum_{a=1}^{k-2}\sum_{b=a}^{k-2} n_{ab}\gamma(\hat{\theta}_{ab}) + \tfrac{1}{2}\sum_{a=1}^{k-2}(n_{a(k-1)}+n_{ak})\gamma(\hat{\theta}_{a(k-1)}) \\
& + \tfrac{1}{2}\sum_{a=1}^{k-1}\sum_{b=k-1}^{k} n_{ab}\gamma(\hat{\theta}_{(k-1)(k-1)}) - \tfrac{1}{2}\sum_{a=1}^{k}\sum_{b=1}^{k}(m_{ab}\log\tfrac{\theta^*_{ab}}{1-\theta^*_{ab}} + n_{ab}\log(1-\theta^*_{ab}))) \\
= \; & \tfrac{1}{2}n^{-1}\sum_{(a,b)\in\mathcal{I}}(m_{ab}\log\tfrac{\theta'_{u(a)u(b)}(1-\theta^*_{ab})}{(1-\theta'_{u(a)u(b)})\theta^*_{ab}} + n_{ab}\log\tfrac{1-\theta'_{u(a)u(b)}}{1-\theta^*_{ab}}) + o_p(1).
\end{aligned}$$

Note that μ can be written as

$$\mu = \frac{1}{2n^2}\sum_{(a,b)\in\mathcal{I}} n_{ab}(\theta^*_{ab}\log\frac{\theta'_{u(a)u(b)}(1-\theta^*_{ab})}{(1-\theta'_{u(a)u(b)})\theta^*_{ab}} + \log\frac{1-\theta'_{u(a)u(b)}}{1-\theta^*_{ab}}).$$

It is easy to see that the expectation of this term is $n\mu$. Therefore, we have

$$\begin{aligned}
& n^{-1}(\max_{z\in[k-1]^n}\sup_{\theta\in\Theta_{k-1}}\log f(A|\theta,z) - \log f(A|\theta^*,z^*)) - n\mu \\
= \; & \tfrac{1}{2}n^{-1}\sum_{(a,b)\in\mathcal{I}}(m_{ab}\log\tfrac{\theta'_{u(a)u(b)}(1-\theta^*_{ab})}{(1-\theta'_{u(a)u(b)})\theta^*_{ab}} + n_{ab}\log\tfrac{1-\theta'_{u(a)u(b)}}{1-\theta^*_{ab}}) - n\mu + o_p(1) \\
= \; & \tfrac{1}{2}n^{-1}\sum_{(a,b)\in\mathcal{I}}(m_{ab} - E(m_{ab})\log\tfrac{\theta'_{u(a)u(b)}(1-\theta^*_{ab})}{(1-\theta'_{u(a)u(b)})\theta^*_{ab}} + o_p(1) \\
\xrightarrow{d} \; & N(0,\sigma(\theta^*)).
\end{aligned}$$

□

8.2 Proofs for Theorem 2

We first prove one useful lemma below.

Lemma 4. *Suppose that $A \sim P_{\theta^*,z^*}$. With probability tending to 1,*

$$\max_{z\in[k]^n}\sup_{\theta\in\Theta_k}\log f(A|\theta,z) = \sup_{\theta\in\Theta_k}\log f(A|\theta,z^*).$$

Proof. This lemma is essentially Lemma 3 in Bickel et al. (2013). The arguments are similar to those for Lemma 2.3 in Wang and Bickel (2016). Note that $G(R(z),\theta^*)$ is maximized at z^*. Define

$$I_{\delta_n} = \{z\in[k]^n : G(R(z),\theta^*) - G(R(z^*),\theta^*) < -\delta_n\},$$

for $\delta_n \to 0$ slowly enough.

By Lemma 2, for $\epsilon_n \to 0$ slowly,

$$\begin{aligned} & \mid F(m(z)/n^2, n(z)/n^2) - G(R(z), \theta^*) \mid \\ \leq\ & C \sum_{a,b=1}^{k} \mid m_{ab}(z)/n^2 - (R\theta^* R^T(z))_{ab} \mid \\ =\ & o_p(\epsilon_n), \end{aligned}$$

since γ is Lipschitz on any interval bounded away from 0 and 1. Then we have

$$\begin{aligned} & \max_{z \in I_{\delta_n}} \sup_{\theta \in \Theta_k} \log f(A|\theta, z) \\ \leq\ & \log(\sum_{z \in I_{\delta_n}} \sup_{\theta \in \Theta_k} f(A|\theta, z)) \\ =\ & \log(\sum_{z \in I_{\delta_n}} \sup_{\theta \in \Theta_k} e^{\log f(A|\theta, z)}) \\ \leq\ & \log(\sup_{\theta \in \Theta_k} f(A|\theta, z^*) k^n e^{o_p(n^2 \epsilon_n) - n^2 \delta_n}) \\ <\ & \sup_{\theta \in \Theta_k} \log f(A|\theta, z^*). \end{aligned}$$

For $z \notin I_{\delta_n}$, $G(R(z), \theta^*) - G(R(z^*), \theta^*) \to 0$. Let $\bar{z} = \min \mid \tau(z) - z^* \mid$. Since the maximum is unique up to τ, $\parallel R(\bar{z}) - R(z^*) \parallel_\infty \to 0$.

By Lemma 2,

$$\begin{aligned} & P(\max_{z \notin \tau(z^*)} \parallel X(\bar{z}) - X(z^*) \parallel_\infty > \epsilon \mid z - z^* \mid /n) \\ \leq\ & \sum_{m=1}^{n} P(\max_{z: z = \bar{z}, |\bar{z} - z^*| = m} \parallel X(\bar{z}) - X(z^*) \parallel_\infty > \epsilon \frac{m}{n}) \\ \leq\ & \sum_{m=1}^{n} 2(k-1)^{k-1} n^m (k-1)^{m+2} e^{-Cnm} \\ \to\ & 0. \end{aligned}$$

It follows for $\mid z - z^* \mid = m$, $z \notin I_{\delta_n}$,

$$\begin{aligned} \parallel \frac{m(\bar{z})}{n^2} - \frac{m(z^*)}{n^2} \parallel_\infty &= o_p(1) \frac{|z - z^*|}{n} + \parallel R\theta^* R^T(\bar{z}) - R\theta^* R^T(z^*) \parallel_\infty \\ &\geq \frac{m}{n}(C + o_p(1)). \end{aligned}$$

Observe $\parallel \frac{m(z^*)}{n^2} - R\theta^* R^T(z^*) \parallel_\infty = o_p(1)$ by Lemma 2, $\parallel \frac{n(z^*)}{n^2} - R\mathbf{1}\mathbf{1}^T R^T(z^*) \parallel_\infty = o_p(1)$. Note that $F(\cdot, \cdot)$ has continuous derivation in the neighborhood of $(\frac{m(z')}{n^2}, \frac{n(z')}{n^2})$. By Lemma 1,

$$\frac{\partial F((1-\epsilon)\frac{m(z^*)}{n^2} + \epsilon M, (1-\epsilon)\frac{n(z^*)}{n^2} + \epsilon t)}{\partial \epsilon} \mid_{\epsilon = 0^+} < -C < 0$$

for (M, t) in the neighborhood of $(\frac{m(z^*)}{n^2}, \frac{n(z^*)}{n^2})$. Thus,

$$F(\frac{m(\bar{z})}{n^2}, \frac{n(\bar{z})}{n^2}) - F(\frac{m(z^*)}{n^2}, \frac{n(z^*)}{n^2}) \leq -C\frac{m}{n}.$$

Since

$$\begin{aligned} & \sup_{\theta \in \Theta_k} \log f(A|\theta, z) - \sup_{\theta \in \Theta_k} \log f(A|\theta, z^*) \\ \leq\ & n^2 (F(\frac{m(\bar{z})}{n^2}, \frac{n(\bar{z})}{n^2}) - F(\frac{m(z^*)}{n^2}, \frac{n(z^*)}{n^2})) \\ =\ & -Cmn, \end{aligned}$$

we have

$$\begin{aligned}
& \max_{z\notin I_{\delta_n}, z\notin\tau(z^*)} \sup_{\theta\in\Theta_k} \log f(A|\theta,z) \\
\leq\ & \log(\sum_{z\notin I_{\delta_n}, z\notin\tau(z^*)} \sup_{\theta\in\Theta_k} f(A|\theta,z)) \\
\leq\ & \log(\sum_{z\in\tau(z^*)} \sup_{\theta\in\Theta_k} f(A|\theta,z) \sum_{m=1}^{n}(k-1)^m n^m e^{-Cmn}) \\
\leq\ & \log(\max_{z\in\tau(z^*)} \sup_{\theta\in\Theta_k} f(A|\theta,z) \sum_{z\in\tau(z^*)} \sum_{m=1}^{n}(k-1)^m n^m e^{-Cmn}) \\
=\ & \sup_{\theta\in\Theta_k} \log f(A|\theta,z^*) + \log(k^k \sum_{m=1}^{n}(k-1)^m n^m e^{-Cmn}) \\
<\ & \sup_{\theta\in\Theta_k} \log f(A|\theta,z^*).
\end{aligned}$$

□

Proof of Theorem 2. By Taylor's expansion, we have

$$\begin{aligned}
& 2(\max_{z\in[k]^n} \sup_{\theta\in\Theta_k} \log f(A|\theta,z) - \log f(A|\theta^*,z^*)) \\
=\ & 2(\sup_{\theta\in\Theta_k} \log f(A|\theta,z^*) - \log f(A|\theta^*,z^*)) \\
=\ & 2\sum_{1\leq a\leq b\leq k}(m_{ab}\log\tfrac{\hat\theta_{ab}}{\theta^*_{ab}} + (n_{ab}-m_{ab})\log\tfrac{1-\hat\theta_{ab}}{1-\theta^*_{ab}}) \\
=\ & 2\sum_{1\leq a\leq b\leq k}(n_{ab}\hat\theta_{ab}\log\tfrac{\theta^*_{ab}+\hat\theta_{ab}-\theta^*_{ab}}{\theta^*_{ab}} + n_{ab}(1-\hat\theta_{ab})\log\tfrac{1-\theta^*_{ab}+\theta^*_{ab}-\hat\theta_{ab}}{1-\theta^*_{ab}}) \\
=\ & 2\sum_{1\leq a\leq b\leq k}(n_{ab}(\theta^*_{ab}+\Delta_{ab})(\tfrac{\Delta_{ab}}{\theta^*_{ab}} - \tfrac{\Delta^2_{ab}}{2\theta^{*2}_{ab}}) + n_{ab}(1-\theta^*_{ab}-\Delta_{ab})(\tfrac{-\Delta_{ab}}{1-\theta^*_{ab}} - \tfrac{\Delta^2_{ab}}{2(1-\theta^*_{ab})^2})) + O(n_{ab}\Delta^3_{ab}) \\
=\ & 2\sum_{1\leq a\leq b\leq k}(n_{ab}(\Delta_{ab} + \tfrac{\Delta^2_{ab}}{2\theta^*_{ab}}) + n_{ab}(-\Delta_{ab} + \tfrac{\Delta^2_{ab}}{2(1-\theta^*_{ab})})) + O(n_{ab}\Delta^3_{ab}),
\end{aligned}$$

where $\Delta_{ab} = \hat\theta_{ab} - \theta^*_{ab}$. By the proof of Theorem 1, $n_{ab}\Delta^3_{ab} = o_p(1)$. Consequently, we have

$$\begin{aligned}
2L_{k,k} &= 2\sum_{1\leq a\leq b\leq k}(\tfrac{n_{ab}\Delta^2_{ab}}{2\theta^*_{ab}} + \tfrac{n_{ab}\Delta^2_{ab}}{2(1-\theta^*_{ab})}) + o_p(1) \\
&= \sum_{1\leq a\leq b\leq k} \tfrac{n_{ab}(\hat\theta_{ab}-\theta^*_{ab})^2}{\theta^*_{ab}(1-\theta^*_{ab})} + o_p(1),
\end{aligned}$$

which converges in distribution to the Chi-square distribution with $k(k+1)/2$ degrees of freedom by the central limit theory. □

8.3 Proof of Theorem 3

. The idea for the proofs is to embed a k-block model in a larger model by appropriately splitting the labels z^*. Define $\nu_{k^+} = \{z \in [k^+]$: there is at most one nonzero entry in every row of $R(z, z^*)\}$. ν_{k^+} is obtained by splitting of z^* such that every block in z is always a subset of an existing block in z^*.

The following lemma will be used in the proof of Theorem 3.

Lemma 5. *Suppose that $A \sim P_{\theta^*,z^*}$. With probability tending to 1,*

$$\max_{z\in[k^+]^n} \sup_{\theta\in\Theta_{k^+}} \log f(A|\theta,z) \leq \alpha n \log k^+ + \sup_{\theta\in\Theta_k} \log f(A|\theta,z^*).$$

Proof. The arguments are similar to those of Lemma 2.6 in Wang and Bickel (2016). Note that $G(R(z),\theta^*)$ is maximized at any $z \in \nu_{k^+}$ with the value $\sum_{a,b} p_a p_b \gamma(\theta^*_{ab})$. Denote the optimal

$G^* = \sum_{a,b} p_a p_b \gamma(\theta^*_{ab})$ and

$$I^+_{\delta_n} = \{z \in [k^+]^n : G(R(z), \theta^*) - G^* < -\delta_n\},$$

for $\delta_n \to 0$ slowly enough.

By Lemma 2, for $\epsilon_n \to 0$ slowly,

$$\mid F(m(z)/n^2, n(z)/n^2) - G(R(z), \theta^*) \mid \leq C \sum_{a,b=1}^{k^+} \mid m_{ab}(z)/n^2 - (R\theta^* R^T(z))_{ab} \mid = o_p(\epsilon_n)$$

since γ is Lipschitz on any interval bounded away from 0 and 1.

It follows from the definition of ν_{k^+} there exists a surjective funciton $h : [k^+] \to [k]$ describing the block assignments in $R(z, z^*)$. For any $z' \in \nu_{k^+}$, it is easy to see

$$\begin{aligned}
& \max_{z \in I^+_{\delta_n}} \sup_{\theta \in \Theta_{k^+}} \log f(A|\theta, z) \\
\leq\ & \log(\textstyle\sum_{z \in I^+_{\delta_n}} \sup_{\theta \in \Theta_{k^+}} f(A|\theta, z)) \\
=\ & \log(\textstyle\sum_{z \in I^+_{\delta_n}} \sup_{\theta \in \Theta_{k^+}} e^{\log f(A|\theta, z)}) \\
\leq\ & \log(\sup_{\theta \in \Theta_{k^+}} f(A|\theta, z')(k^+ - 1)^n e^{o_p(n^2\epsilon_n) - n^2\delta_n}) \\
<\ & \log(\sup_{\theta \in \Theta_{k^+}} f(A|\theta, z')) \\
=\ & \sup_{\theta \in \Theta_{k^+}} \log f(A|\theta, z') \\
=\ & \sup_{\theta \in \Theta_{k^+}} \tfrac{1}{2} \textstyle\sum_{a,b=1}^{k} \sum_{(u,v) \in h^{-1}(a) \times h^{-1}(b)} (m_{uv} \log \theta_{uv} + (n_{uv} - m_{uv}) \log(1 - \theta_{uv})).
\end{aligned}$$

Let

$$L_{ab} = \sum_{(u,v) \in h^{-1}(a) \times h^{-1}(b)} (m_{uv} \log \theta_{uv} + (n_{uv} - m_{uv}) \log(1 - \theta_{uv}) + \lambda(\sum_{(u,v) \in h^{-1}(a) \times h^{-1}(b)} n_{uv} - n_{ab}).$$

Let

$$\frac{\partial L_{ab}}{\partial n_{uv}} = \log(1 - \theta_{uv}) + \lambda = 0.$$

This implies that for $(u, v) \in h^{-1}(a) \times h^{-1}(b)$, θ_{uv}'s are all equal. Let $\theta_{uv} = \theta_{ab}$. Hence,

$$\begin{aligned}
& \textstyle\sum_{(u,v) \in h^{-1}(a) \times h^{-1}(b)} (m_{uv} \log \theta_{uv} + (n_{uv} - m_{uv}) \log(1 - \theta_{uv}) \\
=\ & m_{ab} \log \theta_{ab} + (n_{ab} - m_{ab}) \log(1 - \theta_{ab}),
\end{aligned}$$

where $m_{ab} = \sum_{(u,v) \in h^{-1}(a) \times h^{-1}(b)} m_{uv}$ and $n_{ab} = \sum_{(u,v) \in h^{-1}(a) \times h^{-1}(b)} n_{uv}$.

Thus,

$$\begin{aligned}
& \max_{z \in I^+_{\delta_n}} \sup_{\theta \in \Theta_{k^+}} \log f(A|\theta, z) \\
\leq\ & \sup_{\theta \in \Theta_{k^+}} \tfrac{1}{2} \textstyle\sum_{a,b=1}^{k} \sum_{(u,v) \in h^{-1}(a) \times h^{-1}(b)} (m_{uv} \log \theta_{uv} + (n_{uv} - m_{uv}) \log(1 - \theta_{uv})) \\
=\ & \sup_{\theta \in \Theta_k} \tfrac{1}{2} \textstyle\sum_{a,b=1}^{k} (m_{ab} \log \theta_{ab} + (n_{ab} - m_{ab}) \log(1 - \theta_{ab})) \\
=\ & \sup_{\theta \in \Theta_k} \log f(A|\theta, z^*).
\end{aligned}$$

Note that treating $R(z)$ as a vector. For every $z \notin I^*_{\delta_n}$, $z \notin \nu_{k^+}$, let $z_\perp$ be such that

$R(z_\perp) = \min_{R(z'):z'\in\nu_{k^+}} \parallel R(z) - R(z') \parallel_2$. $R(z) - R(z')$ is perpendicular to the corresponding $k^+ - k$ face. This orthogonality implies the directional derivative of $G(\cdot, \theta^*)$ along the direction of $R(z) - R(z')$ is bounded away from 0. That is

$$\frac{\partial G((1-\epsilon)R(z_\perp) + \epsilon R(z), \theta^*)}{\partial \epsilon} \mid_{\epsilon=0^+} < -C$$

for some universal positive constant C. Similar to the proof in Lemma 3,

$$\sup_{\theta\in\Theta_{k^+}} \log f(A|\theta, z) - \sup_{\theta\in\Theta_{k^+}} \log f(A|\theta, z_\perp) \le -Cmn,$$

where $\mid z - z_\perp \mid = m$. For some $0 < \alpha \le 1 - \frac{C}{\log k^+} + \frac{2\log n + \log k}{n \log k^+}$, we have

$$\begin{aligned}
& \max_{z\notin I^+_{\delta_n}, z\notin\nu_{k^+}} \sup_{\theta\in\Theta_{k^+}} \log f(A|\theta, z) \\
\le\ & \log(\textstyle\sum_{z\notin I^+_{\delta_n}, z\notin\nu_{k^+}} \sup_{\theta\in\Theta_{k^+}} f(A|\theta, z)) \\
\le\ & \log(\textstyle\sum_{z\in\nu_{k^+}} \sup_{\theta\in\Theta_{k^+}} f(A|\theta, z) \sum_{m=1}^n (k-1)^m n^m e^{-Cnm}) \\
\le\ & \log(\textstyle\sum_{z\in\nu_{k^+}} \sup_{\theta\in\Theta_{k^+}} f(A|\theta, z) \sum_{m=1}^n (k-1)^m n^m e^{-Cnm}) \\
\le\ & \log \mid \nu_{k^+} \mid + \max_{z\in\nu_{k^+}} \sup_{\theta\in\Theta_{k^+}} \log f(A|\theta, z) + \log(\textstyle\sum_{m=1}^n (k-1)^m n^m e^{-Cnm}) \\
<\ & \log \mid \nu_{k^+} \mid + \max_{z\in\nu_{k^+}} \sup_{\theta\in\Theta_{k^+}} \log f(A|\theta, z) + \log(n^2 k e^{-Cn}) \\
\le\ & n \log k^+ + \max_{z\in\nu_{k^+}} \sup_{\theta\in\Theta_{k^+}} \log f(A|\theta, z) + 2\log n + \log k - Cn \\
\le\ & \alpha n \log k^+ + \sup_{\theta\in\Theta_k} \log f(A|\theta, z^*).
\end{aligned}$$

□

Proof of Theorem 3. By Lemma 5 and Theorem 2,

$$\begin{aligned}
& \max_{z\in[k^+]^n} \sup_{\theta\in\Theta_{k^+}} \log f(A|\theta, z) - \log f(A|\theta^*, z^*) \\
\le\ & \alpha n \log k^+ + \sup_{\theta\in\Theta_k} \log f(A|\theta, z^*) - \log f(A|\theta^*, z^*) \\
=\ & \alpha n \log k^+ + \tfrac{1}{2} \textstyle\sum_{1\le a\le b\le k} \frac{n_{ab}(\hat{\theta}_{ab} - \theta^*_{ab})^2}{\theta^*_{ab}(1-\theta^*_{ab})} + o_p(1) \\
=\ & \alpha n \log k^+ + O_p(\tfrac{k(k+1)}{2} \log n)
\end{aligned}$$

□

8.4 Proof of Theorem 4

Let

$$g_n(k, \lambda, A) = \max_{z\in[k]^n} \sup_{\theta\in\Theta_k} \log f(A|\theta, z) - (\lambda n \log k + \frac{k(k+1)}{2} \log n),$$

and

$$h_n(k, \lambda, A) = \max_{z\in[k]^n} \sup_{\theta\in\Theta_k} \log f(A|\theta, z) - \log f(A|\theta^*, z^*) - (\lambda n \log k + \frac{k(k+1)}{2} \log n).$$

For $k' < k$, by Theorem 2, we have

$$\begin{aligned}
& P(\ell(k') > \ell(k)) = P(g_n(k', \lambda, A) > g_n(k, \lambda, A)) \\
= \;& P(h_n(k', \lambda, A) > h_n(k, \lambda, A)) \\
= \;& P(h_n(k', \lambda, A) > \sup_{\theta \in \Theta_k} \log f(A|\theta, z^*) - \log f(A|\theta^*, z^*) - (\lambda n \log k + \tfrac{k(k+1)}{2} \log n)) \\
= \;& P(\max_{z \in [k']^n} \sup_{\theta \in \Theta_{k'}} \log f(A|\theta, z) - \log f(A|\theta^*, z^*) > \lambda(n \log k' - n \log k) \\
& \quad + (\tfrac{k'(k'+1)}{2} \log n - \tfrac{k(k+1)}{2} \log n) + \sup_{\theta \in \Theta_k} \log f(A|\theta, z^*) - \log f(A|\theta^*, z^*)) \\
= \;& P(\max_{z \in [k']^n} \sup_{\theta \in \Theta_{k'}} \log f(A|\theta, z) - \log f(A|\theta^*, z^*) > \lambda(n \log k' - n \log k) \\
& \quad + (\tfrac{k'(k'+1)}{2} \log n - \tfrac{k(k+1)}{2} \log n) + O_p(\tfrac{k(k+1)}{2} \log n)) \\
= \;& P(n^{-1}(\max_{z \in [k']^n} \sup_{\theta \in \Theta_{k'}} \log f(A|\theta, z) - \log f(A|\theta^*, z^*)) - n\mu \\
& \quad > -n\mu + n^{-1}(\lambda(n \log k' - n \log k) + (\tfrac{k'(k'+1)}{2} \log n - \tfrac{k(k+1)}{2} \log n)) + O_p(\tfrac{k(k+1)}{2n} \log n)).
\end{aligned}$$

By Theorem 1, the above probability goes to zero.

For $k' > k$, we have

$$\begin{aligned}
& P(\ell(k') > \ell(k)) = P(g_n(k', \lambda, A) > g_n(k, \lambda, A)) \\
= \;& P(h_n(k', \lambda, A) > h_n(k, \lambda, A))) \\
\leq \;& P(\alpha n \log k' + \sup_{\theta \in \Theta_k} \log f(A|\theta, z^*) - \log f(A|\theta^*, z^*) - (\lambda n \log k' + \tfrac{k'(k'+1)}{2} \log n) > \\
& \sup_{\theta \in \Theta_k} \log f(A|\theta, z^*) - \log f(A|\theta^*, z^*) - (\lambda n \log k + \tfrac{k(k+1)}{2} \log n)).
\end{aligned}$$

By Theorem 3, for $\lambda > (\alpha \log k')/(\log k' - \log k)$, the above probability goes to zero.

8.5 Proof of Theorem 7

For $a = k-1$, $b = k-1$, $u(a) = k-1$, we have

$$\begin{aligned}
\tfrac{\theta'_{u(a)u(b)}}{\theta^*_{ab}} &= \tfrac{\theta'_{(k-1)(k-1)}}{\theta^*_{(k-1)(k-1)}} = \tfrac{\hat{\theta}_{(k-1)(k-1)}}{\theta^*_{(k-1)(k-1)}} + o_p(\tfrac{\log n}{n}) \\
&= \tfrac{m_{(k-1)(k-1)} + m_{(k-1)k}}{n_{(k-1)(k-1)} + n_{(k-1)k}} / \theta^*_{(k-1)(k-1)} + o_p(\tfrac{\log n}{n}) \\
&= \tfrac{n_{(k-1)(k-1)}\theta^*_{(k-1)(k-1)} + n_{(k-1)k}\theta^*_{(k-1)k}}{n_{(k-1)(k-1)} + n_{(k-1)k}} / \theta^*_{(k-1)(k-1)} + o_p(\tfrac{\log n}{n}) \\
&= x_1 + (1 - x_1) \tfrac{\theta^*_{(k-1)k}}{\theta^*_{(k-1)(k-1)}} + o_p(\tfrac{\log n}{n}) \\
&= x_1 + \tfrac{q}{p}(1 - x_1) + o_p(\tfrac{\log n}{n}),
\end{aligned}$$

where $x_1 = \frac{n_{(k-1)(k-1)}}{n_{(k-1)(k-1)} + n_{(k-1)k}}$. It yields

$$\theta^*_{(k-1)(k-1)} \log \frac{\theta'_{(k-1)(k-1)}}{\theta^*_{(k-1)(k-1)}} = p \log(x_1 + \frac{q}{p}(1 - x_1)) + o(1).$$

Again, we have

$$\begin{aligned}
\frac{1-\theta'_{u(a)u(b)}}{1-\theta^*_{ab}} &= \frac{1-\theta'_{(k-1)(k-1)}}{1-\theta^*_{(k-1)(k-1)}} = \frac{1-\hat{\theta}_{(k-1)(k-1)}}{1-\theta^*_{(k-1)(k-1)}} + o_p(\frac{\log n}{n}) \\
&= (1 - \frac{m_{(k-1)(k-1)}+m_{(k-1)k}}{n_{(k-1)(k-1)}+n_{(k-1)k}})/(1-\theta^*_{(k-1)(k-1)}) + o_p(\frac{\log n}{n}) \\
&= \frac{n_{(k-1)(k-1)}(1-\theta^*_{(k-1)(k-1)})+n_{(k-1)k}(1-\theta^*_{(k-1)k})}{n_{(k-1)(k-1)}+n_{(k-1)k}}/(1-\theta^*_{(k-1)(k-1)}) + o_p(\frac{\log n}{n}) \\
&= x_1 + \frac{1-q}{1-p}(1-x_1) + o_p(\frac{\log n}{n}).
\end{aligned}$$

It yields

$$(1-\theta^*_{(k-1)(k-1)}) \log \frac{1-\theta'_{(k-1)(k-1)}}{1-\theta^*_{(k-1)(k-1)}} = (1-p)\log(x_1 + \frac{1-q}{1-p}(1-x_1)) + o(1).$$

For $a = k$, $b = k-1$, $u(a) = k-1$, we have

$$\begin{aligned}
\frac{\theta'_{u(a)u(b)}}{\theta^*_{ab}} &= \frac{\theta'_{(k-1)(k-1)}}{\theta^*_{k(k-1)}} = \frac{\hat{\theta}_{(k-1)(k-1)}}{\theta^*_{k(k-1)}} + o_p(\frac{\log n}{n}) \\
&= \frac{m_{(k-1)(k-1)}+m_{(k-1)k}}{n_{(k-1)(k-1)}+n_{(k-1)k}}/\theta^*_{k(k-1)} + o_p(\frac{\log n}{n}) \\
&= \frac{n_{(k-1)(k-1)}\theta^*_{(k-1)(k-1)}+n_{(k-1)k}\theta^*_{(k-1)k}}{n_{(k-1)(k-1)}+n_{(k-1)k}}/\theta^*_{k(k-1)} + o_p(\frac{\log n}{n}) \\
&= \frac{p}{q}x_1 + (1-x_1) + o_p(\frac{\log n}{n}).
\end{aligned}$$

It yields

$$\theta^*_{k(k-1)} \log \frac{\theta'_{(k-1)(k-1)}}{\theta^*_{k(k-1)}} = q\log(\frac{p}{q}x_1 + (1-x_1)) + o(1).$$

Again, we have

$$\begin{aligned}
\frac{1-\theta'_{u(a)u(b)}}{1-\theta^*_{ab}} &= \frac{1-\theta'_{(k-1)(k-1)}}{1-\theta^*_{k(k-1)}} = \frac{1-\hat{\theta}_{(k-1)(k-1)}}{1-\theta^*_{k(k-1)}} + o_p(\frac{\log n}{n}) \\
&= (1 - \frac{m_{(k-1)(k-1)}+m_{(k-1)k}}{n_{(k-1)(k-1)}+n_{(k-1)k}})/(1-\theta^*_{k(k-1)}) + o_p(\frac{\log n}{n}) \\
&= \frac{n_{(k-1)(k-1)}(1-\theta^*_{(k-1)(k-1)})+n_{(k-1)k}(1-\theta^*_{(k-1)k})}{n_{(k-1)(k-1)}+n_{(k-1)k}}/(1-\theta^*_{k(k-1)}) + o_p(\frac{\log n}{n}) \\
&= \frac{1-p}{1-q}x_1 + (1-x_1) + o_p(\frac{\log n}{n}).
\end{aligned}$$

It yields

$$(1-\theta^*_{k(k-1)}) \log \frac{1-\theta'_{(k-1)(k-1)}}{1-\theta^*_{k(k-1)}} = (1-q)\log(\frac{1-p}{1-q}x_1 + (1-x_1)) + o(1).$$

Let $s = p/q$ and $\beta_1 = \frac{1-x_1}{sx_1}$. Noting that $x_1 + \frac{q}{p}(1-x_1) = x_1(1+\beta_1)$, we have

$$p\log(x_1 + \frac{q}{p}(1-x_1)) = p\log x_1 + p\log(1+\beta_1).$$

Under the condition that $s = p/q$ is sufficiently large, compared with $p\log x_1$, $p\log(1+\beta_1)$ is very small and can be omitted .

By similar discussions,

$$(1-p)\log(x_1+\frac{1-q}{1-p}(1-x_1))=(1-p)\log\frac{1-px_1}{1-p}+(1-p)\log(1+\beta_2),$$

$$q\log(\frac{p}{q}x_1+(1-x_1))=q\log(\frac{px_1}{q})+q\log(1+\beta_3),$$

$$(1-q)\log(\frac{1-p}{1-q}x_1+(1-x_1))=(1-q)\log(1-px_1))+(1-q)\log(1+\beta_4).$$

Thus,

$$\begin{aligned} n\mu &= \tfrac{nC_3}{2}(p\log x_1+(1-p)\log\tfrac{1-px_1}{1-p}+q\log(\tfrac{px_1}{q})+(1-q)\log(1-px_1))+I_5+o(n)\\ &\triangleq I_1+I_2+I_3+I_4+I_5+o(n).\\ &= O(n)+O(n)+npx_1(\tfrac{q}{px_1}\log(\tfrac{px_1}{q}))+O(n)+I_5+o(n), \end{aligned}$$

where $I_5=\frac{nC_3}{2}(p\log(1+\beta_1)+(1-p)\log(1+\beta_2)+q\log(1+\beta_3)+(1-q)\log(1+\beta_4))$.

Compared with $I_i<0$, $i=1,2,3,4$, I_5 is very small and can be omitted. Note that $I_1<0$, $I_2<0$, $I_3>0$ and $I_4<0$. For large p/q, $\frac{q}{px_1}\log(\frac{px_1}{q})$ is very small. That is, $I_3<\mid I_1\mid$. Thus, we get $n\mu\to-\infty$.

8.6 Proof of Theorem 8

By Taylor expansion, we have

$$\begin{aligned} & 2(\max_{z\in[k]^n}\sup_{\theta\in\Theta_k}\log f(A|\theta,\omega,z)-\log f(A|\theta^*,\omega,z^*))\\ =\ & 2(\sup_{\theta\in\Theta_k}\log f(A|\theta,\omega,z^*)-\log f(A|\theta^*,\omega,z^*))\\ =\ & 2\sum_{1\le a\le b\le k}(m_{ab}\log\tfrac{\hat\theta_{ab}}{\theta^*_{ab}}-n_{ab}(\hat\theta_{ab}-\theta^*_{ab}))\\ =\ & 2\sum_{1\le a\le b\le k}(n_{ab}\hat\theta_{ab}\log\tfrac{\theta^*_{ab}+\hat\theta_{ab}-\theta^*_{ab}}{\theta^*_{ab}})-n_{ab}(\hat\theta_{ab}-\theta^*_{ab})\\ =\ & 2\sum_{1\le a\le b\le k}(n_{ab}(\theta^*_{ab}+\Delta_{ab})(\tfrac{\Delta_{ab}}{\theta^*_{ab}}-\tfrac{\Delta^2_{ab}}{2\theta^{*2}_{ab}})-n_{ab}\Delta_{ab}+O(n_{ab}\Delta^3_{ab})\\ =\ & 2\sum_{1\le a\le b\le k}(\tfrac{n_{ab}\Delta^2_{ab}}{\theta^*_{ab}}-\tfrac{n_{ab}\Delta^2_{ab}}{2\theta^*_{ab}})+O(n_{ab}\Delta^3_{ab})\\ =\ & \sum_{1\le a\le b\le k}\tfrac{n_{ab}\Delta^2_{ab}}{\theta^*_{ab}}+O(n_{ab}\Delta^3_{ab})\\ =\ & \sum_{1\le a\le b\le k}\tfrac{n_{ab}(\hat\theta_{ab}-\theta^*_{ab})^2}{\theta^*_{ab}}+o(1) \end{aligned}$$

where $\Delta_{ab}=\hat\theta_{ab}-\theta^*_{ab}$. Since

$$\sum_{1\le a\le b\le k}\frac{n_{ab}(\hat\theta_{ab}-\theta^*_{ab})^2}{\theta^*_{ab}}\le\max_{1\le a\le b\le k}(1-\theta^*_{ab})\sum_{1\le a\le b\le k}\frac{n_{ab}(\hat\theta_{ab}-\theta^*_{ab})^2}{\theta^*_{ab}(1-\theta^*_{ab})}$$

and

$$\sum_{1\le a\le b\le k}\frac{n_{ab}(\hat\theta_{ab}-\theta^*_{ab})^2}{\theta^*_{ab}}\ge\min_{1\le a\le b\le k}(1-\theta^*_{ab})\sum_{1\le a\le b\le k}\frac{n_{ab}(\hat\theta_{ab}-\theta^*_{ab})^2}{\theta^*_{ab}(1-\theta^*_{ab})},$$

we have

$$\sum_{1\le a\le b\le k} \frac{n_{ab}(\hat{\theta}_{ab}-\theta^*_{ab})^2}{\theta^*_{ab}} = O_p(\frac{k(k+1)}{2}\log n).$$

Acknowledgements

Hu's work was partly done while he was visiting Professor Harrison Zhou at the Yale University during 2015–2016. The authors sincerely thank Harrison Zhou for helpful discussions and valuable suggestions.

References

Adamic, L.A. and Glance, N. (2005). The political blogosphere and the 2004 US election. *In Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem 36-43.* ACM.

Airoldi, E. M., Blei, D. M., Finberg, S. E. and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**, 1981-2014.

Amini, A. A., Chen, A., Bickel, P. J. and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41**, 2097-2122.

Barabási, A. L. and Bonabau, E. (2003). Scale-free networks, *Scientific American*, 50–59.

Bickel, P. J. and Chen, A. (2009). A nonparameter view of network models and Newman-Girvan and other modularities. *Proc. Nat. Acad. Sci.* **106**, 21068-21073.

Bickel, P. J. and Sarkar, P. (2015). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B*, **78**, 253-273.

Bickel, P. J., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41**, 1922-1943.

Celisse, A., Daudin, J. J., Pierre, L. (2012). Consistency of maximum likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* **6**, 1847-1899.

Chen, K. and Lei, J. (2017). Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, to appear.

Choi, D. S., Wolfe, P. J. and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99**, 273-284.

Daudin, J. J., Picard, F. and Robin, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18**, 173-183.

Decelle, A., Krzakala, F., Moore, C. and Zdeborova, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**, 066106.

Fishkind, D. E., Sussman, D. L., Tang, M., Vogelstein, J. T. and Priebe, C. E. (2013). Consistent adjacency-spectral partioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34**, 23-39.

Gao, C., Lu, Y., and Zhou, H. H. (2015). Rate optimal graphon estimation. *Ann. Statist.* **43**, 2624-2652.

Gao, C., Ma, Z., Zhang, A. Y. and Zhou, H. H. (2016). Community detection in degree corrected block models. *arXiv preprint arXiv: 1607.06993*.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13–30.

Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5**, 109-137.

Jin, J. (2015). Fast network community detection by SCORE. *Ann. Statist.* **43**, 57-89.

Joseph, A. and Yu, B. (2016). Impact of regularization on spectral clustering. *Ann. Statist.* **44**, 1765-1791.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks *Phys. Rev. E* **83**, 016107.

Latouche, P. Birmele, E. and Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Stat. Modelling* **12**, 93-115.

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

Lei J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.*, **44**, 401–424.

Newman, M. E. J. (2006a). Modularity and community structure in networks. *Proc. Nat. Acad. Sci.* **103**, 8577-8582.

Newman, M. E. J. (2006b). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, **96**, 1077–1087.

Peixoto, T. P. (2013). Parsimonious module inference in large networks. *Phys. Rev. Lett.* **110**, 148701.

Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic block model. *Ann. Statist.* **39**, 1878-1915.

Saldana, D. F., Yu, Y. and Feng, Y. (2017). How many communities are there?. *Journal of Computational and Graphical Statistics*, **26**, 171-181.

Snijders, T. A. B. and Nowicki, K. (1997). Estimation and prediction for stochastic block models for graphs with latent block structure. *Journal of Classification*, **14**, 75–100.

Wang, R., and Bickel, P. J. (2016). Likelihood-based model selection for stochastic block models. *Ann. Statist.*, to appear.

Westveld, A. H. and Hoff, P. D. (2011). A mixed effects model for longitudinal relational and network data with applications to international trade and conflict. *The Annals of Applied Statistics*, 5, 843–872.

Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. *arXiv preprint arXiv: 1309.5936*.

Zhang, A. Y. and Zhou, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.*, **44**, 2252-2280.

Zhao, Y., Levina, E., and Zhu, J. (2011). Community extraction for social networks. *Proc. Nat. Acad. Sci.* **108**, 7321-7326.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under

degree-corrected stochastic block models. *Ann. Statist.* **40**, 2266-2292.