LIKELIHOOD-BASED MODEL SELECTION FOR STOCHASTIC BLOCK MODELS∗

BY Y.X. RACHEL WANG† AND PETER J. BICKEL‡

Department of Statistics, University of California, Berkeley‡

The stochastic block model (SBM) provides a popular framework for modeling community structures in networks. However, more attention has been devoted to problems concerning estimating the latent node labels and the model parameters than the issue of choosing the number of blocks. We consider an approach based on the log likelihood ratio statistic and analyze its asymptotic properties under model misspecification. We show the limiting distribution of the statistic in the case of underfitting is normal and obtain its convergence rate in the case of overfitting. These conclusions remain valid in the dense and semi-sparse regimes. The results enable us to derive the correct order of the penalty term for model complexity and arrive at a likelihood-based model selection criterion that is asymptotically consistent. In practice, the likelihood function can be estimated by more computationally efficient variational methods, allowing the criterion to be applied to moderately large networks.

1. Introduction. Network modeling has attracted increasing research attention in the past few decades as the amount of data on complex systems accumulates at an unprecedented rate. Many complex systems in science and nature consist of interacting individual components which can be represented as nodes with connecting edges in a network. Network modeling has found numerous applications in studying friendship networks in sociology, Internet traffic in information technology, predator-prey interactions in ecology, and protein-protein interactions and gene regulatory mechanisms in molecular biology.

One prominent feature of many of these networks is the presence of communities, where groups of nodes exhibit high internal connectivity. Communities provide a natural division of the network into subunits with certain traits. In social networks, they often arise based on people’s common interests and geographic locations. The World Wide Web forms communities or hubs based on the content of the web pages. In gene networks, communities correspond to genes with related functional groupings, many of which
can act in the same biological pathway. Numerous heuristic algorithms have been proposed for detecting communities. However, a generative model is needed to study the problem from a theoretical perspective.

The stochastic block model (SBM), proposed by Holland, Laskey and Leinhardt (1983) in social science, is one of the simplest random graph models incorporating community structures. It assigns each node a latent discrete block variable and the connectivity levels between nodes are determined by their block memberships. In practice, this model sometimes oversimplifies the structures of real networks and other variants have been proposed, including the degree-corrected SBM (Karrer and Newman, 2011) relaxing the within-block degree homogeneity constraint and overlapping SBM (Airoldi et al., 2008) allowing a node to be in multiple blocks. These models have been applied to model real networks in social science and biology (Bickel and Chen, 2009; Daudin, Picard and Robin, 2008; Airoldi et al., 2008; Karrer and Newman, 2011).

Much research effort has been devoted to the problems of estimating the latent block memberships and model parameters of a SBM, including modularity (Newman, 2006a) and likelihood maximization (Bickel and Chen, 2009; 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), belief propagation (Decelle et al., 2011) to name but a few. The asymptotic properties of some of these methods have also been studied (Bickel and Chen, 2009; Rohe, Chatterjee and Yu, 2011; Celisse et al., 2012; Bickel et al., 2013). However, these methods require knowing (or knowing at least a suitable range for) \( K \), the number of blocks, a priori. Less attention has been paid to the problem of selecting \( K \). For general networks this corresponds to the issue of determining the number of communities, which remains a challenging open problem. Recursive approaches have been adopted to extract (Zhao, Levina and Zhu, 2011) or partition (Bickel and Sarkar, 2013) one community sequentially, while using optimization strategies or hypothesis testing to decide whether the process should be stopped at one stage. A more general sequential test for comparing a fitted SBM against alternative models with finer structures is proposed in Lei (2014). Conceptually these approaches are more appealing for networks with a hierarchical structure. In other cases, it would be more desirable to be able to compare different community numbers directly. A few likelihood-based model selection criteria have been proposed (Daudin, Picard and Robin, 2008; Latouche, Birmele and Ambroise, 2012; Saldana, Yu and Feng, 2014). From an information-theoretic perspective, Peixoto (2013) proposed a criterion based on minimum length description. These ap-
proaches circumvent the difficulty of analyzing the likelihood directly by using variational approximations or assuming the node labels are fixed and using plug-in estimates obtained from other inference algorithms. Furthermore, the asymptotic studies of these criteria examining their large-sample performance remain incomplete. Empirically, a network cross-validation method has been investigated in Chen and Lei (2014).

In this paper, we directly address the challenges involved in analyzing the asymptotic distribution of the maximum log likelihood function under model misspecification. We show the log likelihood ratio statistic is asymptotically normal in the case of underfitting. Although obtaining an explicit asymptotic distribution of the statistic in the case of overfitting is much more challenging, we have still derived its order of convergence and subsequently shown these two cases of misspecification can be separated with probability tending to one. We thus propose a model selection criterion taking the form of a penalized likelihood and show it is asymptotically consistent. Our conclusions remain valid for networks with average degree growing at a poly-log rate in the semi-sparse regime. Computationally the likelihood can be approximated with the variational algorithm in Daudin, Picard and Robin (2008), making this approach applicable to reasonably large networks. We also provide comparisons of its performance on simulated and real networks with other model selection approaches.

2. Results.

2.1. Preliminaries. A SBM with $K$ blocks on $n$ nodes is defined as follows. A vector of latent labels $Z = (Z_1, \ldots, Z_n)$ is generated with $Z_i$ taking integer values from $[K] = \{1, \ldots, K\}$ governed by a multinomial distribution with parameters $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$. Given $Z_i = a, Z_j = b$, an adjacency matrix $A$ is generated with

$$A_{i,j} | (Z_i = a, Z_j = b) \sim \text{Bernoulli}(H_{a,b}), \quad i \neq j.$$ 

We consider a symmetric $A$ with zero diagonal entries corresponding to an undirected graph, although our arguments generalize easily to directed graphs. $H$ is a $K \times K$ symmetric matrix describing the connectivities within and between blocks. We denote the model parameters $\theta = (\pi, H)$ and let $\Theta_K$ be the parameter space of a $K$-block model,

$$\Theta_K = \{ \theta \mid \pi \in (0, 1)^K, \sum_{a=1}^K \pi_a = 1, H \in (0, 1)^{K \times K} \}.$$
Throughout the paper, $\theta^* = (\pi^*, H^*)$ will denote the true generative parameter giving rise to an observed $A$. We will further parametrize $H^*$ by $H^* = \rho_n S^*$, where the degree density $\rho_n$ may be $\Omega(1)$ or going to zero at a rate $n \rho_n / \log n \to \infty$. We assume $\theta^* \in \Theta_K$ and $H^*$ has no identical columns, meaning the underlying model has $K$ blocks and it is identifiable in the sense that it cannot be further collapsed to a smaller model.

$z = (z_1, \ldots, z_n) \in [K']^n$ represents another set of labels under a $K'$-block model with $K'$ not necessarily equaling $K$. $g(A; \theta)$ is the likelihood function describing the distribution of $A$ with parameter $\theta \in \Theta_{K'}$ and can be written as the sum of the complete likelihood function $f(z, A; \theta)$ associated with the labels $z \in [K']^n$:

$$g(A; \theta) = \sum_{z \in [K']^n} f(z, A; \theta),$$

where $f(z, A; \theta)$ takes the form

$$f(z, A; \theta) = \left( \prod_{i=1}^{n} \pi_{z_i} \right) \left( \prod_{i<j} H_{z_i, z_j}^{A_{i,j}} (1 - H_{z_i, z_j})^{1-A_{i,j}} \right)$$

$$= \left( \prod_{a=1}^{K'} \pi_a^{n_a(z)} \right) \left( \prod_{a=1}^{K'} \prod_{b=1}^{K'} H_{a,b}^{O_{a,b}(z)} (1 - H_{a,b})^{n_{a,b}(z) - O_{a,b}(z)} \right)^{1/2}$$

with count statistics

$$n_a(z) = \sum_{i=1}^{n} \mathbb{I}(z_i = a), \quad n_{a,b}(z) = \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}(z_i = a, z_j = b)$$

$$O_{a,b}(z) = \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}(z_i = a, z_j = b) A_{i,j}.$$

$g$ and $f$ are invariant with respect to a permutation on the block labels, $\tau: [K'] \to [K']$, and its corresponding permutations on the node labels $z$ and the parameters $\theta$. Furthermore, let $R(z)$ be the $K' \times K$ confusion matrix whose $(k, a)$-th entry is

$$R_{k,a}(z, Z) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(z_i = k, Z_i = a).$$

We take a likelihood-based approach toward model selection and first investigate whether different model choices can be separated using the log
likelihood ratio

\[ L_{K,K'} = \log \frac{\sup_{\theta \in \Theta_{K'}} g(A; \theta)}{\sup_{\theta \in \Theta_K} g(A; \theta)}. \]

Here the comparison is made between the correct \( K \)-block model and fitting a misspecified \( K' \)-block model.

In the following sections, we analyze the asymptotic distribution of \( L_{K,K'} \) for \( K' \neq K \). The main focus of analysis lies in handling the sum in (2.1) which contains an exponential number of terms. It has been shown in Bickel et al. (2013) that when \( \theta \in \Theta_K \), \( \sup_{\theta \in \Theta_K} g(A; \theta) \) is essentially equivalent to maximizing the complete likelihood corresponding to the correct labels \( Z \), \( \sup_{\theta \in \Theta_K} f(Z,A; \theta) \). In the next section, we first show an analogous result in the case of underfitting and use it to derive the asymptotic distribution of \( L_{K,K'} \).

2.2. Underfitting. We start by considering \( K' = K - 1 \). Intuitively, a \((K-1)\)-block model can be obtained by merging blocks in a \( K \)-block model. More specifically, given the correct 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 \( U_{a,b}(H^*, p) \) which combines blocks \( a \) and \( b \) in \( H^* \) by taking weighted averages with proportions in \( p \). For example, for \( H = U_{K-1,K}(H^*, p) \),

\[
\begin{align*}
H_{l,k} &= H_{l,k}^* \quad \text{for } 1 \leq l, k \leq K - 2; \\
H_{l,K-1} &= \frac{p_l p_{K-1} H_{l,K-1}^* + p_k p_{K} H_{K-1,K}^*}{p_l p_{K-1} + p_k p_{K}} \quad \text{for } 1 \leq l \leq K - 2; \\
H_{K-1,K-1} &= \frac{p_{K-1}^2 H_{K-1,K-1}^* + 2p_{K-1} p_{K} H_{K-1,K}^* + p_{K}^2 H_{K,K}^*}{p_{K-1}^2 + 2p_{K-1} p_{K} + p_{K}^2}.
\end{align*}
\]

A schematic representation of \( H \) is given in Figure 1.

For consistency, when merging two blocks \((a,b)\) with \( b > a \), the new merged block will be relabeled \( a \) and all the blocks \( c \) with \( c > b \) will be relabeled \( c - 1 \). Using this scheme, we also obtain the merged node labels \( U_{a,b}(Z) \) and merged proportions \( U_{a,b}(p) \) with \( [U_{a,b}(p)]_a = p_a + p_b \).

Constraining the parameters to a smaller model results in a suboptimal likelihood and its distance from the likelihood associated with the correct model can be measured by the Kullback-Leibler divergence, denoted \( D_{KL}(\cdot \| \cdot) \). Let

\[
\gamma_1(x) = x \log x + (1 - x) \log(1 - x),
\gamma_2(x) = x \log x - x.
\]
Fig 1: A schematic representation of how $H^*$ is merged to give $H = U_{K-1,K}(H^*, p)$. The green area contains unchanged parameters and the arrows indicate where mergings occur.

and define

\begin{equation}
D_i(a, b) = \sum_{k, l=1}^{K-1} [U_{a,b}(\pi^*)]_k [U_{a,b}(\pi^*)]_l \gamma_i([U_{a,b}(H^*, \pi^*)]_{k,l})
\end{equation}

When $p = \pi^*$ and treating the labels $Z$ as fixed parameters, denote $P_{A|Z,H^*}$ the probability distribution of $A$. Then the information loss incurred by the merging operation $U_{a,b}$ can be measured by

\begin{equation}
D_{KL} \left( P_{A|Z,H^*} || P_{A|U_{a,b}(Z), U_{a,b}(H^*, \pi^*)} \right)
\end{equation}

Thus an optimal merging minimizing $D_{KL}$ is essentially equivalent to maximizing $D_i(a, b)$.

We assume the following holds for $\theta^*$:

**Assumption 2.1.** A unique maximum exists for $\max_{(a,b)} D_i(a, b)$.

This assumption is more of a notational convenience than necessity. From now on without loss of generality assume the maximum is achieved at $a = K - 1$ and $b = K$, and denote $H' = U_{K-1,K}(H^*, \pi^*)$, $S' = H'/\rho_n$ and $Z' = U_{K-1,K}(Z)$. We also assume $H'$ is identifiable in the sense that
Assumption 2.2. \( H' \) has no identical columns.

Thus the merged model cannot be collapsed further to a smaller model.

The next lemma argues \( \sup_{\theta \in \Theta_{K-1}} g(A; \theta) \) is essentially dominated by the complete likelihood associated with the optimal merging.

Lemma 2.3. Let \( S(z) \) be the set of labels which are equivalent up to a permutation \( \tau \), \( S(z) = \{ \tau(z), \tau : [K-1] \rightarrow [K-1] \} \). Then

\[
\sum_{z \notin S(Z')} \sup_{\theta \in \Theta_{K-1}} f(z, A; \theta) = \sup_{\theta \in \Theta_{K-1}} f(Z', A; \theta) o_P(1).
\]

The proof is shown in the Appendix.

This lemma provides a tractable bound on \( \sup_{\theta \in \Theta_{K-1}} g(A; \theta) \), allowing the rest of the analysis to be carried out by usual Taylor expansion. Define

\[
\mu_1(\theta^*) = \frac{1}{2} \left[ D_1(K-1, K) - \sum_{c,d=1}^K \pi_{c,d}^* \gamma_1(H_{c,d}^*) \right]
\]

\[
\mu_2(\theta^*) = \mu_1 + \frac{1}{n} \left\{ (\pi_{K-1}^* + \pi_K^*) \log(\pi_{K-1}^* + \pi_K^*) - \pi_{K-1}^* \log \pi_{K-1}^* - \pi_K^* \log \pi_K^* \right\}
\]

The following theorem gives the asymptotic distribution of \( L_{K,K-1} \), the proof of which is shown in the Appendix.

Theorem 2.4. Suppose the underlying model parameter generating \( A \) is \( \theta^* = (\pi^*, H^*) \in \Theta_K \), then \( L_{K,K-1} \) is asymptotically normal with

\[
n^{-3/2} L_{K,K-1} - \sqrt{n} \mu_1(\theta^*) \xrightarrow{D} N(0, \sigma_1^2(\theta^*)), \text{ if } \rho_n = \Omega(1);
\]

\[
\rho_n^{-1} n^{-3/2} L_{K,K-1} - \rho_n^{-1} \sqrt{n} \mu_2(\theta^*) \xrightarrow{D} N(0, \sigma_2^2(\theta^*)), \text{ if } \rho_n \rightarrow 0.
\]

Let \( I \) be the set of indices affected by the merge \( U_{K-1,K}(H^*, \pi^*) \),

\[ I = \{(a,b) \in [K]^2 \mid K-1 \leq a \leq K \text{ or } K-1 \leq b \leq K \}, \]

and \( u(a) \) such that

\[
u(a) = \begin{cases} a & \text{for } a \leq K - 2 \\ K-1 & \text{for } K - 1 \leq a \leq K. \end{cases}
\]

Define \( d_i = (d_i(a,b))_{(a,b) \in I, a \leq b} \) as

\[
d_1(a,b) = H_{a,b}^* \log \frac{H'_{u(a),K-1}^*}{H_{a,b}^*} + (1 - H_{a,b}^*) \log \frac{1 - H'_{u(a),K-1}^*}{1 - H_{a,b}^*}
\]
\[ d_2(a, b) = S_{a,b}^* \log \frac{S'_{a(a), K-1}}{S_{a,b}^*} + (S'_{a(a), K-1} - S_{a,b}^*). \]

Denote \( \Sigma(\pi^*) \) the covariance matrix of a multinomial(\( \pi^* \)) distribution, \( B(x) \) the Jacobi matrix of the vector valued function \( \xi(x_1, \ldots, x_K) = (\xi_{a,b})_{(a,b) \in I, a \leq b} \), where

\[
\xi_{a,b} = \begin{cases} 
  x_a x_b & \text{for } a \neq b \\
  \frac{x^2}{2} & \text{for } a = b.
\end{cases}
\]

The variance \( \sigma_i(\theta^*) \) is given by \( d_i^T B(\pi^*) \Sigma(\pi^*) B(\pi^*)^T d_i \) for \( i = 1, 2 \).

**Remark 2.5.**

(i) In general, underfitting a \( K^- < K \) model will lead to the same type of limiting distribution under conditions similar to Assumptions 2.1 and 2.2, assuming the uniqueness of the optimal merging scheme and identifiability after merging. That is,

\[
\rho_n^{-1} n^{-3/2} L_{K,K^-} - \rho_n^{-1} \sqrt{n} \mu \xrightarrow{D} N(0, \sigma^2)
\]

for some mean \( \mu \sim C \rho_n \) and variance \( \sigma^2 \). The proof will be similar but involve more tedious descriptions of how various merges can occur.

(ii) The asymptotic distributions derived under the null distribution of a \( K \)-block model suggest one might consider performing hypothesis testing directly to compare against an alternative simpler model. However, the asymptotic means depend on the true parameters, and its maximum likelihood estimate converges only at the rate \( \sqrt{n} \) (Bickel et al., 2013).

(iii) Without Assumptions 2.1 and 2.2, it is easy to show

\[
L_{K,K^-} \leq -\Omega_P(n^2 \rho_n),
\]

where \( \Omega(\cdot) \) denotes asymptotic lower bound, using the method in proving Theorem 2.7.

### 2.3. Overfitting.

In the case of overfitting a \( K^+ > K \) model, deriving the asymptotic distribution of \( L_{K,K^+} \) is much more challenging. Intuitively, 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. We first show a result analogous to Lemma 2.3. However, the number of summands involved in \( \sup_{\theta \in \Theta_{K^+}} g(A; \theta) \) remains exponential this time.

Recall that for \( z \in [K^+]^n \), \( R(z, Z) \) is the \( K^+ \times K \) confusion matrix. We first define a subset \( \mathcal{V}_{K^+} \in [K^+]^n \) such that

\[
\mathcal{V}_{K^+} = \{ z \in [K^+]^n \mid \text{there is at most one nonzero entry in every row of } R(z, Z) \}.
\]
$V_{K^+}$ is obtained by splitting of $Z$ such that every block in $z$ is always a subset of an existing block in $Z$. The next lemma shows it suffices to consider only the subclass of labels $V_{K^+}$ in the sum $g(A; \theta)$, the proof of which is given in the Appendix.

**Lemma 2.6.** Suppose $\theta^* \in \Theta_K$, then

$$\sum_{z \in [K^+]^n} \sup_{\theta \in \Theta_{K^+}} f(z, A; \theta) = (1 + o_P(1)) \sum_{z \in V_{K^+}} \sup_{\theta \in \Theta_{K^+}} f(z, A; \theta).$$

The lemma does not provide a direct simplification of the sum and suggests the reason why obtaining an asymptotic distribution for $L_{K,K^+}$ is difficult. On the other hand, with appropriate concentration we can still derive the asymptotic order of the statistic.

**Theorem 2.7.** Suppose $\theta^* \in \Theta_K$, then overfitting by a $K^+$-block model with $K^+ > K$ gives $L_{K,K^+} = O_P(n^{3/2} \rho_n^{1/2})$.

The proof is provided in the Appendix.

2.4. Model selection. The results in the previous sections lead us to construct a penalized likelihood criterion for selecting the optimal block number. The criterion is consistent in the sense that asymptotically it chooses the correct $K$ with probability one. Define

$$\beta(K') = \sup_{\theta \in \Theta_{K'}} \log g(A; \theta) - N_{K'} B_n,$$

where $B_n$ gives the order of the penalty term, and $N_{K'}$ is a strictly increasing sequence indexed by $K'$ describing the complexity of the model. The optimal $K_0$ is such that

$$K_0 = \arg \max_{K'} \beta(K').$$

**Corollary 2.8.** For $K' < K$, setting $B_n = o(n^2 \rho_n)$,

$$\mathbb{P}_{\theta^*}(\beta(K') < \beta(K)) \to 1. \quad (2.13)$$

For $K' > K$, setting $B_n$ such that $B_n n^{-3/2} \rho_n^{-1/2} \to \infty$,

$$\mathbb{P}_{\theta^*}(\beta(K') < \beta(K)) \to 1. \quad (2.14)$$
Proof. For $K' < K$, generalizing Theorem 2.4,

$$
\Pr_{\theta^*} \left( \beta(K') < \beta(K) \right) \\
= \Pr_{\theta^*} \left( -\frac{n^{-3/2} \rho_n^{-1} \log \sup_{\theta \in \Theta_{K'}} g(A; \theta)}{\sup_{\theta \in \Theta_K} g(A; \theta)} - \sqrt{n \rho_n^{-1} \mu} < (N_{K'} - N_K) \frac{B_n}{n^{3/2} \rho_n^{1/2}} - \sqrt{n \rho_n^{-1} \mu} \right) \\
\to 1,
$$

(2.15)

since $B_n = o(n^2 \rho_n)$ and $-\rho_n^{-1} \mu \geq C(\theta^*)$ for some positive constant depending on $\theta^*$. In general the same conclusion holds by Remark 2.5 (iii).

For $K' > K$, using Theorem 2.7,

$$
\Pr_{\theta^*} \left( \beta(K') < \beta(K) \right) \\
= \Pr_{\theta^*} \left( -\frac{1}{n^{3/2} \rho_n^{1/2} \rho_n} \log \sup_{\theta \in \Theta_{K'}} g(A; \theta) < (N_{K'} - N_K) \frac{B_n}{n^{3/2} \rho_n^{1/2}} \right) \\
\to 1,
$$

(2.16)

when $B_n n^{-3/2} \rho_n^{-1/2} \to \infty$.

Since the ratio of the upper bound $n^2 \rho_n$ and the lower bound $n^{3/2} \rho_n^{1/2}$ tends to infinity, such a sequence $B_n$ exists. Choosing $B_n$ in this interval, we have $K_0 = K$ with probability tending to 1. However, we also note that for finite cases with moderate-sized $n$, $\sqrt{n \rho_n^{-1} \mu}$ in (2.15) is small, making it easy to over penalize with large $B_n$. At the same time, the lower bound in Theorem 2.7 is not tight and can be refined further.

We further assume the followings hold for tractable approximation.

**Assumption 2.9.**

$$
\sum_{z \in V_{K+}} \sup_{\theta \in \Theta_{K+}} f(z, A; \theta) = O_P(e^{M_n}),\text{ where}
$$

(2.17)

$$
M_n = \max_{z \in V_{K+}} \sup_{\theta \in \Theta_{K+}} \log f(z, A; \theta)
$$

**Assumption 2.10.** The maximum is achieved in the set $N_{K+} = \{z \in V_{K+} \mid n_k(z) \geq \epsilon n \text{ for all } k, \text{ for some } \epsilon > 0, \}$.

Assumption 2.9 assumes a Laplace-type approximation holds for the sum, whereas Assumption 2.10 assumes the maximum can only be achieved on a loosely balanced block design. These assumptions together with Lemma 2.6
imply it remains to analyze the order of \( \max_{z \in \mathcal{N}_{K^+}} \sup_{\theta \in \Theta_{K^+}} \log f(z, A; \theta) \). The following theorem shows the order of \( L_{K,K^+} \) can be refined to \( O_P(1) \). The details can be found in the Appendix.

**Theorem 2.11.** Under Assumptions 2.9 and 2.10, \( L_{K,K^+} \) is of order \( O_P(1) \) for \( K^+ > K \).

It follows then choosing \( B_n \) growing slightly faster than a constant will ensure consistency in the sense described in Corollary 2.8. It can also be deduced from the proof that the order of \( L_{K,K^+} \) grows at most at the rate \((K^++1)K^+/2\). Thus we choose a penalized likelihood of the following form,

\[
\beta(K') = \sup_{\theta \in \Theta_{K'}} \log g(A; \theta) - \lambda \cdot \frac{K'(K'+1)}{2} \log n,
\]

where the constant \( \lambda \) is a tuning parameter and does not affect the asymptotic properties of the criterion. It is not surprising that the penalty term has the same order as other BIC-type criteria (e.g. Daudin, Picard and Robin (2008)) based on the complete likelihood assuming the node labels are fixed. Recall that in the underfitting case we have proved the likelihood is essentially equivalent to the complete likelihood corresponding to the appropriate labels. A similar equivalence also holds for the overfitting case by Lemma 2.6 and Assumption 2.9.

2.5. Approximation by variational likelihood. In practice, direct computations of the likelihood function \( g(A; \theta) \) involves an exponential number of summands and quickly become intractable as \( n \) grows. In particular, the optimization over \( \theta \) using the EM algorithm requires computing the conditional distribution of \( Z \) given \( A \), which is not factorizable in this case. Variational methods tackle the true conditional distribution \( f_{Z|A,\theta} \) with the mean field approximation, thus simplifying the local optimization at each iteration. The variational log likelihood \( J(q, \theta; A) \) for a \( K' \)-block model is defined as

\[
J(q, \theta; A) = -D_{KL}(q || f_{Z|A,\theta}) + \log g(A; \theta),
\]

where \( q \in \mathcal{D}_{K'} \) is any product distribution with \( q(z) = \prod_{i=1}^{n} q_i(z_i) \), \( 1 \leq z_i \leq K' \). The variational estimates \( \hat{\theta}^{\text{VAR}}_{K'} \) is given by

\[
\hat{\theta}^{\text{VAR}}_{K'} = \arg \max_{\theta \in \Theta_{K'}} \max_{q \in \mathcal{D}_{K'}} J(q, \theta; A),
\]
which can be optimized using the EM algorithm in Daudin, Picard and Robin (2008). Also we note that $J(q, \theta; A)$ simplifies to

$$J(q, \theta; A) = \sum_{i=1}^{n} \sum_{k=1}^{K'} q_i(k)(-\log q_i(k) + \log \pi(k))$$

$$+ \sum_{i<j} \sum_{k,l=1}^{K'} q_i(k)q_j(l) (A_{ij} \log H_{k,l} + (1 - A_{ij}) \log(1 - H_{k,l}))$$

and hence can be easily evaluated.

We can replace the likelihood in (2.18) by the variational log likelihood $J$ without changing its asymptotic performance. More precisely, the criterion with variational approximation

$$(2.20) \quad \beta_{\text{VAR}}(K') = \sup_{\theta \in \Theta_{K'}} \sup_{q \in D_{K'}} J(q, \theta; A) - \lambda \cdot \frac{K'(K' + 1)}{2} \log n$$

is still asymptotically consistent. Noting that

(i) $\sup_{\theta \in \Theta_{K'}} \sup_{q \in D_{K'}} J(q, \theta; A) \leq \sup_{\theta \in \Theta_{K'}} \log g(A; \theta)$;
(ii) $\sup_{\theta \in \Theta_{K'}} \sup_{q \in D_{K'}} J(q, \theta; A) - \sup_{\theta \in \Theta_{K'}} \log g(A; \theta) = O_P(1)$ as shown in Bickel et al. (2013),

it can be easily verified that (2.15) and (2.16) still hold.

3. Simulations. We first examined how well the normal limiting distribution approximated the empirical distribution of the statistic in the case of underfitting. Figure 2 plots the distribution of $n^{-3/2}L_{K,K-1}$ for $n = 200$ and $n = 500$ obtained from 200 replications for the following two scenarios:

(a) $K = 2$, $\pi^* = (0.4, 0.6)$, $H^* = \left( \begin{array}{ccc} 0.15 & 0.05 & 0.01 \\ 0.2 & 0.1 & 0.1 \\ 0.2 & 0.03 \end{array} \right)$;  
(b) $K = 3$, $\pi^* = (0.4, 0.3, 0.3)$, $H^* = \left( \begin{array}{ccc} 0.2 & 0.1 & 0.1 \\ 0.2 & 0.03 \\ 0.1 \end{array} \right)$.

The log likelihoods are approximated by the variational EM algorithm initialized by regularized spectral clustering (Joseph and Yu, 2013). The solid curves are normal densities with mean $\mu_2(\theta^*)$ and $\sigma(\theta^*)$ given in Theorem 2.4. Even though the $O(n)$ term in $\mu_2(\theta^*)$ diminishes asymptotically for $\rho_n$ going to 0 slowly, we found it essential to correct for the bias in the finite sample regimes above. In both cases, the convergence to the Gaussian shape appears faster than the convergence to the mean, and a bias exists.
for \( n = 200 \). When the network size reaches 500, the empirical distributions are well approximated by their limiting distribution. We note that the bias should not have an adverse effect on model selection since it is in the direction away from zero, making it easier to separate the two models.

Fig 2: Empirical distributions of \( n^{-3/2} L_{K,K-1} \) for (a), (b) \( K=2 \), \( \pi^* \) and \( H^* \) as described in scenario (a); (c), (d) \( K=3 \), \( \pi^* \) and \( H^* \) as described in scenario (b). \( n = 200 \) in (a) and (c); \( n = 500 \) in (b) and (d). The solid curves are normal densities with mean \( \mu_2(\theta^*) \) and \( \sigma(\theta^*) \) as given in Theorem 2.4.
Next we investigated how the success rate of the criterion (2.20) changes with respect to the tuning parameter $\lambda$. Figure 3 shows the fraction of the penalized likelihood selecting the correct $K$ out of 50 trials for $\lambda$ values varying between 0 and 4. The generative parameters for $K = 2$ and $K = 3$ are given in scenarios (a) and (b), and in addition a $K = 5$ model was generated with $\pi^*_i = 0.2$ for all $i$ and the entries in $H^*$ varying between 0.06 and 0.19. For $K = 2$ and $K = 3$, the penalized likelihood achieves reasonable success rate for $\lambda$ smaller than 3 when the network size reaches 200. When $n = 500$, the success rate appears robust to the choice of $\lambda$ and is maintained at 1 for a wide range of values. For $K = 5$, however, it becomes difficult to select the correct $K$ since the task of fitting also becomes harder as $K$ increases.

![Fig 3: The fraction of the penalized likelihood with different values of $\lambda$ successfully choosing the correct $K$ out of 50 iterations for (a) $K = 2$, $\pi$ and $H^*$ as described in scenario (a); (b) $K = 3$, $\pi$ and $H^*$ as described in scenario (b); (c) $K = 5$, $\pi_i = 0.2$ for all $i$, $H^*$ with entries varying between 0.06 and 0.19.](image-url)

To see how our criterion (denoted $vlh$) compares with other existing model selection methods, we fix $\lambda = 1$ and compare its success rate with variational Bayes (Latouche, Birmele and Ambroise (2012), denoted $vb$) and the 3-fold network cross validation method in Chen and Lei (2014) (denoted $ncv$). In Figure 4, these methods were implemented on 50 networks of size 500 with $K = 2, 3, 4$, $H^* = \rho S^*$, and $\rho \in \{0.02, 0.04, \ldots, 0.1\}$. The average degrees of these networks range from around 12 to 75. In general, the success rate of each method decreases as the networks become sparser and the number of blocks grows. Overall $vlh$ outperforms the other two methods, and although not explicitly shown, the trends remain true for $\lambda$ values between...
0.25 and 2.

Fig 4: Comparison of the success rates of the penalized likelihood ($\lambda = 1$, \(vlh\)) with variational Bayes (\(vb\)) and network cross validation (\(ncv\)) when (a) \(K = 2, \pi = (0.4, 0.6)\); (b) \(K = 3, \pi = (0.3, 0.3, 0.4)\); (c) \(K = 4, \pi_i = 0.25\) for all \(i\). In all the cases, \(H^* = \rho S^*\), where \(\rho \in \{0.02, 0.04, \ldots, 0.1\}\), the diagonal elements of \(S^*\) equal 2 and the off diagonal elements equal 1.

4. Real world networks. We first implemented our method along with \(vb\) and \(ncv\) on nine Facebook ego networks, collected and labeled by Leskovec and Mcauley (2012). An ego network is created by extracting subgraphs formed on the neighbors of a central (ego) node. Any isolated node was removed before analysis. The actual sizes of the networks and the number of communities selected by the three methods are shown in Table 1. The second row of the table shows the number of friend circles in every network with some individuals belonging to multiple circles, but not every individual possesses a circle label. These circle numbers give partial truth on how many communities there are in the networks. Overall, the penalized likelihood and \(vb\) tend to produce comparable community numbers, whereas \(ncv\) consistently favors small community numbers. The penalized likelihood approach is reasonably robust to the choice of \(\lambda\) on larger networks. Both the penalized likelihood and \(vb\) tend to underestimate the community number on smaller networks with a large number of circles and a small average degree, reflecting the difficulty in fitting a large \(K\)-block model on small networks.

We also experimented these methods on the political book network (Newman, 2006b), which consists of 105 books and their edges representing copurchase information from Amazon. Figure 5 (a) shows the manual labeling
of the books based on their political orientations being either conservative, liberal or neutral. (b) and (c) show the community structures obtained by our method with three choices of $\lambda$. When $\lambda = 2$, the method selected $K = 3$ with the clustering of the nodes being close to the truth. With the other two smaller $\lambda$ values, the method selected $K = 6$ and the clustering further splits each of the communities obtained previously into two. vb found four communities but merged two clusters in (a) into one. ncv again produced the smallest $K$ value with $K = 2$.

![Communities in 105 political books](image)

**Table 1**

| # Non-isolated vertices | 333 | 1034 | 224 | 150 | 168 | 61 | 786 | 534 | 52 |
|-------------------------|-----|------|-----|-----|-----|----|-----|-----|----|
| # Circles               | 24  | 9    | 14  | 7   | 13  | 13 | 17  | 32  | 52 |
| Average degree          | 15  | 52   | 29  | 23  | 20  | 9  | 36  | 18  | 6  |
| Optimal $K$, ($\lambda = 1/4$) | 13  | 15   | 15  | 10  | 13  | 9  | 20  | 14  | 6  |
| ($\lambda = 1/2$)       | 13  | 15   | 15  | 10  | 10  | 9  | 20  | 14  | 6  |
| ($\lambda = 1$)         | 10  | 15   | 13  | 7   | 9   | 6  | 20  | 14  | 3  |
| Optimal $K$, vb         | 11  | 24   | 16  | 9   | 11  | 6  | 25  | 23  | 6  |
| Optimal $K$, ncv        | 3   | 6    | 4   | 2   | 4   | 2  | 2   | 2   | 3  |

*Facebook ego networks and the number of communities selected by the three methods, the penalized likelihood with three choices of $\lambda$.*

5. Discussion. In this paper, we have studied the problem of selecting the community number under a regular stochastic block model, allowing the average degree to grow at a polylog rate and the true block number being fixed. Using techniques similar to Bickel et al. (2013), we have shown the log likelihood ratio statistic has an asymptotic normal distribution when a smaller model with fewer blocks is specified. In the case of misfitting a
larger model, we have obtained the convergence rate for the statistic. Combining these results we arrive at a likelihood-based model selection criterion that is asymptotically consistent. For finite-sized networks, we have further refined the bound for the statistic in the overfitting case under reasonable assumptions to correct for the possibility of over-penalizing.

There are a number of open problems for future work. (i) It would be desirable to have a data-driven approach to select the tuning parameter $\lambda$. Similar to other AIC and BIC-type criteria under standard models, the choice of this constant does not affect the asymptotic consistency of the criterion. Our analysis and simulation suggest small $\lambda$ is often preferred to avoid over-penalizing. However, it is less clear on the real networks which $\lambda$ is optimal. (ii) It would be interesting to investigate whether the results can be extended to other block model variants, such as degree-corrected SBM (Karrer and Newman, 2011) and overlapping SBM (Airoldi et al., 2008). (iii) We have performed our analysis with fixed block number as the number of nodes tends to infinity. However, in practice the number of communities is also likely to grow as a network expands (Choi, Wolfe and Airoldi, 2012), especially when we view block models as histogram approximations for more general models (Bickel and Chen, 2009; Wolfe and Olhede, 2013). Peixoto (2013) has provided some analysis on the maximum number of blocks detectable for a given SBM graph with fixed labels. In general as more time-course network data become available in biology, social science, and many other domains, incorporating dynamic features of community structures into network modeling will remain an interesting direction to explore.

**APPENDIX A: PROOFS OF LEMMAS AND THEOREMS**

In this section, we prove all the lemmas and theorems in the main paper. Denote $\mu_{n} = n^2 \rho_{n}$, the total number of edges $L = \sum_{i=1}^{n} \sum_{j=i+1}^{n} A_{i,j}$, and $N(z) = (n_{k,l}(z))_{1 \leq k,l \leq K'}$. For two sets of labels $z$ and $y$, $|z - y| = \sum_{i=1}^{n} \mathbb{I}(z_{i} \neq y_{i})$. $\| \cdot \|_{\infty}$ denotes the maximum norm of a matrix. We abbreviate $R(z,Z)S^{*}R(z,Z)^{T}$ as $RS^{*}R^{T}(z)$. $C, C_1, C_2, \ldots$ are constants which might be different at each occurrence. The following concentration inequalities bound the variations in $A$ and will be used throughout the section.

**Lemma A.1.** Suppose $z \in [K']^{n}$ and define $X(z) = O(z)/\mu_{n} - RS^{*}R^{T}(z)$. For $\epsilon \leq 3$,

$$\mathbb{P}\left( \max_{z \in [K']^{n}} \|X(z)\|_{\infty} \geq \epsilon \right) \leq 2(K')^{n+2} \exp\left( -\frac{1}{4(\|S^{*}\|_{\infty} + 1)} \epsilon^2 \mu_{n} \right).$$
Let \( y \in [K']^n \) be a fixed set of labels, then for \( \epsilon \leq 3m/n \),
\[
\mathbb{P} \left( \max_{z \mid |z-y| \leq \rho} \| X(z) - X(y) \|_\infty > \frac{m}{n} \right) \leq 2 \left( \frac{n}{m} \right) (K')^{m+2} \exp \left( - \frac{n \epsilon^2 \mu_n}{4m(4\|S^*\|_\infty + 1)} \right).
\]  
(A.2)

Proof. The proof follows from Bickel et al. (2013) with minor modifications for general \( K' \)-block models and correcting for the zero diagonal in \( A \). 

Recall that
\[
\gamma_1(x) = x \log x + (1 - x) \log(1 - x),
\]
\[
\gamma_2(x) = x \log x - x.
\]

Define \( F_i(M, t), i = 1, 2 \), as
\[
F_i(M, t) = \sum_{k,l=1}^{K'} t_{k,l} \gamma_i \left( \frac{M_{k,l}}{t_{k,l}} \right),
\]  
(A.3)

Then the log of the complete likelihood can be expressed as
\[
\sup_{\theta \in \Theta_{K'}} \log f(z, A; \theta) = n \sum_{k=1}^{K'} \alpha(n_k(z)/n) + \frac{n^2}{2} F_1 \left( O(z)/n^2, N(z)/n^2 \right),
\]  
(A.4)

where \( \alpha(x) = x \log(x) \). Noting the first term is of smaller order compared to the second term, and the conditional expectation of the argument in \( \gamma_1 \) given \( Z \) is \( [RH^*R^T(z)]_{k,l}/[R11^T R^T(z)]_{k,l} \) and \( [RS^* R^T(z)]_{k,l}/[R11^T R^T(z)]_{k,l} \) for \( \gamma_2 \) (up to a diagonal difference) with fluctuation bounded by Lemma A.1, we will focus on analyzing the conditional expectation
\[
G_1(R(z), H^*) = \sum_{k,l=1}^{K'} [R11^T R^T(z)]_{k,l} \gamma_1 \left( \frac{[RH^*R^T(z)]_{k,l}}{[R11^T R^T(z)]_{k,l}} \right) \quad \text{for } \rho_n = \Omega(1),
\]  
(A.5)

\[
G_2(R(z), H^*) = \sum_{k,l=1}^{K'} ([R11^T R^T(z)]_{k,l} \gamma_2 \left( \frac{[RS^* R^T(z)]_{k,l}}{[R11^T R^T(z)]_{k,l}} \right) \quad \text{for } \rho_n \to 0.
\]  
(A.6)
The following lemma shows in the case of underfitting a \((K-1)\)-block model, to maximize \(G_i\) over different configurations of \(R(z, Z)\) with given \(Z\), it suffices to consider the merging scheme described in Section 2.2 by combining two existing blocks in \(Z\).

**Lemma A.2.** Given the true labels \(Z\) with block proportions \(p = n(Z)/n\), maximizing the function \(G_1(R(z), H^*)\) over \(R\) achieves its maximum in the label set
\[
\{ z \in [K - 1]^n \mid \text{there exists } \tau \text{ such that } \tau(z) = U_{a,b}(Z), 1 \leq a < b \leq K, \}
\]
where \(U_{a,b}\) merges \(Z_i\) with labels \(a\) and \(b\).

Furthermore, suppose \(z_0\) gives the unique maximum (up to permutation \(\tau\)), for all \(R\) such that \(R \geq 0, R^T 1 = p\),
\[
\frac{\partial G_1((1 - \epsilon)R(z_0) + \epsilon R, H^*)}{\partial \epsilon}\bigg|_{\epsilon = 0^+} < -C < 0
\]
for \(\rho_n = \Omega(1)\). The same conclusions hold for \(G_2(R(z), S^*)\).

**Proof.** Treating \(R\) as a \((K - 1) \times K\)-dimensional vector, it is easy to check \(G_1(\cdot, H^*)\) is a convex function. Furthermore, since \(R \geq 0, R^T 1 = p\), the domain is part of a convex polyhedron \(P_R = \{ R \in \mathbb{R}^{K(K - 1)} \mid R \geq 0, R^T 1 = p\}\). Therefore the maximum is attained at the vertices of \(P_R\), that is \(R_{\text{vert}}\) such that for every \(a\), exactly one \(R_{\text{vert}}\) is nonzero. This is equivalent to assigning all \(Z_i \in [K]\) with the same label into one group with a new label in \([K - 1]\). Let \(u : [K] \rightarrow [K - 1]\) be the function specified by \(R_{\text{vert}}\), then
\[
G_1(R_{\text{vert}}, H^*) = \sum_{k,l} \sum_{a \in u^{-1}(k), b \in u^{-1}(l)} p_a p_b \gamma_1 \left( \frac{\sum_{a \in u^{-1}(k), b \in u^{-1}(l)} H_{a,b} p_a p_b}{\sum_{a \in u^{-1}(k), b \in u^{-1}(l)} p_a p_b} \right).
\]

Note that there exists at least one \(l \in [K - 1]\) such that \(|u^{-1}(l)| > 1\), and \(\{u^{-1}(k), k \in [K - 1]\}\) forms a partition on \([K]\). By strict convexity of \(\gamma_1\) and identifiability of \(H^*\), to maximize \(G_1\) it suffices to consider merging two of the labels in \([K]\) and mapping the other labels to the remaining labels in \([K - 1]\) in a one-to-one relationship.

The second part of the lemma holds since it is easy to see when the maximum is unique, the derivative of the \(G_1\) at the optimal vertex is bounded away from 0 in all directions. The same arguments apply to \(G_2\). \(\square\)
Noting that when $p = \pi^*$, $G_i$ evaluated at $R(U_{a,b}(Z))$ is equal to $D_i$ defined in (2.5), it is easy to see Assumptions 2.1 and 2.2 guarantees the maximum is unique. We will now prove Lemma 2.3.

**Proof of Lemma 2.3.** Taking the log of the complete likelihood,

$$
\sup_{\theta \in \Theta_{K-1}} \log f(z, A; \theta) = n \sum_{k=1}^{K-1} \alpha(n_k(z)/n) + \frac{n^2}{2} F_1\left(O(z)/n^2, N(z)/n^2\right).
$$

(A.9)

By concentration of $p_k$, it suffices to consider $\{\|p - \pi^*\|_\infty < \eta\}$, where $\eta$ is small enough that $Z'$ remains the unique maximizer of $G_1(R(z), H^*)$ and $G_2(R(z), S^*)$, and distribution conditional on $Z$.

Using techniques similar to Bickel et al. (2013), we prove this by considering $z$ far away from $Z'$ and close to $Z'$ (up to permutation $\tau$). Let $\delta_n$ be a sequence converging to 0 slowly. Define

$$
I_{\delta_n} = \{z \in [K - 1]^n : G_1(R(z), H^*) - G_1(R(Z'), H^*) < -\delta_n\}.
$$

First by (A.1) in Lemma A.1, for $\epsilon_n \to 0$ slowly,

$$
|F_1\left(O(z)/n^2, N(z)/n^2\right) - G_1(R(z), H^*)| \leq C \sum_{k,l} |O_{k,l}(z)/n^2 - (RH^*R^T(z))_{k,l}| + O(n^{-1})
$$

(A.10)

$$
= o_P(\epsilon_n)
$$

since $\gamma_1$ is Lipschitz on any interval bounded away from 0 and 1 and $\min H^* = \Omega(1)$. For $z \in I_{\delta_n}$ and $\rho_n = \Omega(1),

$$
\sum_{z \in I_{\delta_n}} \sup_{\theta \in \Theta_{K-1}} e^{\log f(z, A; \theta)} \leq \sup_{\theta \in \Theta_{K-1}} f(Z', A; \theta)(K - 1)^n e^{O(n) + o_P(n^2 \epsilon_n) - n^2 \delta_n}
$$

(A.11)

$$
= \sup_{\theta \in \Theta_{K-1}} f(Z', A; \theta) o_P(1)
$$

choosing $\delta_n \to 0$ slowly enough such that $\delta_n/\epsilon_n \to \infty$. Similarly for $\rho_n \to 0$, define

$$
J_{\delta_n} = \{z \in [K - 1]^n : G_2(R(z), S^*) - G_2(R(Z'), S^*) < -\delta_n\}.
$$
Note that in this case, for \( \epsilon_n \to 0 \) slowly,
\[
F_1 \left( O(z)/n^2, N(z)/n^2 \right)
= 2 \log \rho_n L/n^2 + \rho_n F_2 \left( O(z)/\mu_n, N(z)/n^2 \right) + O_P(\rho_n^2)
\]
(A.12)
\[
\leq 2 \log \rho_n L/n^2 + \rho_n G_2 \left( R(z), S^* \right) + o_P(\rho_n \epsilon_n) + O_P(\rho_n^2),
\]
by (A.1) and the fact that \( \gamma_2 \) is Lipschitz on any interval bounded away from 0 and 1 and \( \min S^* > 0 \). Then for \( z \in J_{\delta_n} \),
\[
\sum_{z \in J_{\delta_n}} \sup_{\theta \in \Theta_{K-1}} e^{\log f(z; A; \theta)}
\leq \sup_{\theta \in \Theta_{K-1}} f(z', A; \theta) (K - 1)^n e^{O(n) + o_P(\mu_n \rho_n) + o_P(\mu_n \epsilon_n) - \mu_n \delta_n}
\]
(A.13)
\[
\leq \sup_{\theta \in \Theta_{K-1}} f(z', A; \theta) o_P(1).
\]
choosing \( \epsilon_n \to 0, \delta_n \to 0 \) slowly enough.

For \( z \notin J_{\delta_n}, |G_2(R(z), H^*) - G_2(R(Z'), H^*)| \to 0 \). Let \( \bar{z} = \min_\tau |\tau(z) - Z'| \). Since the maximum is unique up to \( \tau \), \( \|R(\bar{z}) - R(Z')\|_\infty \to 0 \), and \( |\sum_k \alpha(n_k(\bar{z})/n) - \sum_k \alpha(n_k(Z')/n)| \to 0 \).

By (A.2),
\[
P \left( \max_{z \notin \bar{S}(Z')} \|X(z) - X(Z')\|_\infty > \epsilon |\bar{z} - Z'|/n \right)
\leq \sum_{m=1}^n P \left( \max_{z: |z - \bar{z}| = m} \|X(z) - X(Z')\|_\infty > \epsilon m/n \right)
\]
(A.14)
\[
\leq \sum_{m=1}^n 2(K - 1)^{-1} m^{(K-1)n} (K - 1)^{m+2} \exp \left( -\frac{m \mu_n}{n} \right) \to 0.
\]

It follows for \( |\bar{z} - Z'| = m, z \notin J_{\delta_n}, \)
\[
\left\| \frac{O(\bar{z})}{\mu_n} - \frac{O(Z')}{\mu_n} \right\|_\infty = o_P(1) \left| \frac{\bar{z} - Z'}{n} \right| + \|R S^* R T(\bar{z}) - R S^* R T(Z')\|_\infty
\]
(A.15)
\[
\geq \frac{m}{n} (C + o_P(1)).
\]

Observe \( \|O(Z')/\mu_n - R S^* R(Z')\|_\infty = o_P(1) \) by Lemma A.2, \( N(Z'/n^2) = R 11^T R T(Z') + o(1) \) on \{ \|p - \pi^*\|_\infty < \eta \}, and \( F_2(\cdot, \cdot) \) has continuous derivative in the neighborhood of \( (O(Z')/\mu_n, N(Z')/n^2) \). Using (A.7) in Lemma A.2,
\[
\frac{\partial F_2 \left( (1 - \epsilon) \frac{O(Z')}{\mu_n} + \epsilon M, (1 - \epsilon) \frac{N(Z')}{n^2} + \epsilon t \right)}{\partial \epsilon}_{\epsilon = 0^+} < -\Omega_P(1) < 0
\]
for \((M,t)\) in the neighborhood of \((O(Z')/\mu_n, N(Z')/n^2)\). Hence

\[
F_2(O(\bar{z})/\mu_n, N(\bar{z})/n^2) - F_2(O(Z')/\mu_n, N(Z')/n^2) \leq - \Omega_P(1) \frac{m}{n}.
\]

We have

\[
\sup_{\theta \in \Theta} \log f(z, A; \theta) - \sup_{\theta \in \Theta} \log f(Z', A; \theta)
\leq n \sum_{k=1}^{K-1} \alpha(n_k(\bar{z})/n) - \alpha(n_k(Z')/n)
+ n^2 (F_1(O(\bar{z})/\mu_n, N(z)/n^2) - F_1(O(Z')/\mu_n, n(Z')/n^2))
\leq (O(n) + o_P(\mu_n) - \Omega_P(\mu_n)) \frac{m}{n}
= - \Omega_P(\mu_n) \frac{m}{n}
\]

(A.17)

using (A.12) and (A.16). We can conclude

\[
\sum_{z \not\in J_n, z \not\in r(Z')} \sup_{\theta \in \Theta} e^{\log f(z, A; \theta)}
\leq \sup_{\theta \in \Theta} f(Z', A; \theta) \sum_{m=1}^{n} (K-1)^{K-1} n^m (K-1)^m e^{-\Omega(\mu_n) m/n}
\leq \sup_{\theta \in \Theta} f(Z', A; \theta) o_P(1)
\]

(A.18)

The bounds (A.13) and (A.18) yield (2.7). The case for \(\rho_n = \Omega(1)\) can be shown in a similar way. \(\square\)

Now Theorem 2.4 follows by Taylor expansion.

**Proof of Theorem 2.4.** First note that

\[
L_{K,K-1} = \log \sup_{\theta \in \Theta_{K-1}} \frac{g(A; \theta)}{g(A; \theta^*)} - \log \sup_{\theta \in \Theta_K} \frac{g(A; \theta)}{g(A; \theta^*)}
= \sup_{\theta \in \Theta_{K-1}} \log \left[ \frac{g(A; \theta)}{f(Z, A; \theta^*)} \cdot \frac{f(Z, A; \theta^*)}{g(A; \theta^*)} \right] + O_P(1)
\]

(A.19)
by a consequence of Theorem 1 and Lemma 3 in Bickel et al. (2013). Noting that \( \sup_{\theta \in \Theta_{K-1}} f(Z', A; \theta) \) is uniquely maximized at (omitting the argument \( Z \))

\[
\hat{\pi}_a = \frac{n_a}{n} = \pi_a^* + O_P(n^{-1/2}) \quad \text{for} \quad 1 \leq a \leq K - 2, \quad \hat{\pi}_{K-1} = \frac{n_{K-1} + n_K}{n}
\]

\[
\hat{H}_{a,b} = \frac{O_{a,b}}{n_{a,b}} = H_{a,b}^* + O_P(\sqrt{\rho_n}n^{-1}) \quad \text{for} \quad 1 \leq a \leq b \leq K - 2,
\]

\[
\hat{H}_{a,K-1} = \frac{O_{a,K-1} + O_{a,K}}{n_{a,K-1} + n_{a,K}} = H_{a,K-1}^* + O_P(\sqrt{\rho_n}n^{-1}) \quad \text{for} \quad 1 \leq a \leq K - 2,
\]

\[
\hat{H}_{K-1,K-1} = \frac{\sum_{a=K-1}^{K-1} \sum_{b=K-1}^{K} O_{a,b}}{\sum_{a=K-1}^{K-1} \sum_{b=K-1}^{K} n_{a,b}} = H_{K-1,K-1}^* + O_P(\sqrt{\rho_n}n^{-1}),
\]

and Assumption 2.2 the merged \( H' \) is identifiable, we have

\[
\sup_{\theta \in \Theta_{K-1}} \sum_{z \in S(Z')} f(z, A; \theta) = 1 + o_P(1).
\]

Combined with Lemma 2.3

\[
\sup_{\theta \in \Theta_{K-1}} \log \frac{g(A; \theta)}{f(Z, A; \theta^*)} = \sup_{\theta \in \Theta_{K-1}} \log \frac{f(Z', A; \theta)}{f(Z, A; \theta^*)} + o_P(1).
\]

(A.20)

We will check the expansion for the case \( \rho_n \to 0 \); the case \( \rho_n = \Omega(1) \) can be shown in the same way.

\[
n^{-3/2} \rho_n^{-1} \sup_{\theta \in \Theta_{K-1}} \log \frac{g(A; \theta)}{f(Z, A; \theta^*)}
\]

\[
= n^{-3/2} \rho_n^{-1} \sup_{\theta \in \Theta_{K-1}} \log \frac{f(Z', A; \theta)}{f(Z, A; \theta^*)} + o_P(1)
\]

\[
= n^{-3/2} \rho_n^{-1} \left\{ \sum_{a=1}^{K-1} \alpha(a) (\hat{\pi}_a) + \sum_{a=1}^{K-2} \sum_{b=a}^{K-2} n_{a,b} \gamma_1(\hat{H}_{a,b}) + \sum_{a=1}^{K-2} (n_{a,K-1} + n_{a,K}) \gamma_1(\hat{H}_{a,K-1})
\right.
\]

\[
\left. + \frac{1}{2} \sum_{a=K-1}^{K-1} \sum_{b=K-1}^{K-1} n_{a,b} \gamma_1(\hat{H}_{K-1,K-1}) - \sum_{a=1}^{K} n_{a} \log \pi_a^* - \frac{1}{2} \sum_{a=1}^{K} \sum_{b=1}^{K} \left( O_{a,b} \log \frac{H_{a,b}^*}{1 - H_{a,b}^*} + n_{a,b} \log(1 - H_{a,b}^*) \right) \right\} + o_P(1)
\]
\[ \rho \]

\[ \text{Lemma 2.3.} \]

Note that in this case the form of \( \sigma \) where the form of \( \rho \) defines similarly to Lemma 2.3.

It follows using arguments similar to Proof of Lemma 2.6.

It suffices to discuss the case \( \sigma \) with the value \( \rho \to 0 \). It is easy to see the expectation of this term is \( \rho^{-1} \sqrt{n} \mu_2 \), we have

\[ \frac{1}{2n^{3/2} \rho_n} \sum_{(a,b) \in I} \left( (O_{a,b} - \mathbb{E}(O_{a,b})) \log \frac{H'_{u(a),u(b)}(1 - H_{a,b}^*)}{(1 - H'_{u(a),u(b)})H_{a,b}^*} + (n_{a,b} - \mathbb{E}(n_{a,b})) \log \frac{1 - H'_{u(a),u(b)}}{1 - H_{a,b}^*} \right) + o_P(1) \]

\[ \rightarrow N(0, \sigma^2(\theta^*)) \]

where the form of \( \sigma^2(\sigma^*) \) can be checked by Taylor expansion and the delta method.

\[ \square \]

**Proof of Lemma 2.6.** The proof follows using arguments similar to Lemma 2.3. Note that in this case \( G_1(R(z), H^*) \) is maximized at any \( z \in V_{K^+} \) with the value \( \sum_{a,b} p_a p_b \gamma_1(H_{a,b}^*) \) (or \( \sum_{a,b} p_a p_b \gamma_2(S_{a,b}^*) \) for \( G_2(R(z), S^*) \)).

It suffices to discuss the case \( \rho_n \to 0 \). Denote the optimal \( G^* := \sum_{a,b} p_a p_b \gamma_2(S_{a,b}^*) \), define similarly to Lemma 2.3.

\[ J_{\delta_n} = \{ z \in [K^+]^n : G_2(R(z), S^*) - G^* < -\delta_n \} \]

for \( \delta_n \to 0 \) slowly enough. It is easy to see

\[ \sum_{z \in J_{\delta_n}} \sup_{\theta \in \Theta_{K^+}} f(z, A; \theta) \leq \sup_{\theta \in \Theta_{K^+}} f(z_0, A; \theta) o_P(1) \]

for any \( z_0 \in V_{K^+} \).
Next note that treating \( R(z) \) as a vector, \( \{ R(z) \mid z \in V_{K+} \} \) is a subset of the union of some of the \( K^+ - K \) faces of the polyhedron \( P_R \). For every \( z \notin J_{\delta_n}, z \notin V_{K+} \), let \( z_\perp \) be such that \( R(z_\perp) := \min_{R(z_0) : z_0 \in V_{K+}} \| R(z) - R(z_0) \|_2 \). \( R(z) - R(z_\perp) \) is perpendicular to the corresponding \( K^+ - K \) face. Furthermore, this orthogonality implies the directional derivative of \( G_2(\cdot, S^*) \) along the direction of \( R(z) - R(z_\perp) \) is bounded away from 0. That is
\[
\frac{\partial G_2 ((1 - \epsilon)R(z_\perp) + \epsilon R(z), S^*)}{\partial \epsilon} \bigg|_{\epsilon=0^+} < -C
\]
for some universal positive constant \( C \). Similar to (A.17),
\[
\sup_{\theta \in \Theta_{K+}} \log f(z, A; \theta) - \sup_{\theta \in \Theta_{K+}} \log f(z_\perp, A; \theta) \leq -\Omega_P(\mu_n) \frac{m}{n} \quad \sup_{\theta \in \Theta_{K+}} f(z, A; \theta) \leq e^{-\Omega_P(\mu_n) \frac{m}{n}} \sup_{\theta \in \Theta_{K+}} f(z_\perp, A; \theta)
\]
where \( |z - z_\perp| = m \). We have
\[
\sum_{z \notin J_{\delta_n}, z \notin V_{K+}} \sup_{\theta \in \Theta_{K+}} f(z, A; \theta) \\
\leq \sum_{z \in V_{K+}} \sup_{\theta \in \Theta_{K+}} f(z, A; \theta) \sum_{m=1}^{n} (K - 1)^m n^m e^{-\Omega_P(\mu_n) \frac{m}{n}} \\
= o_P(1) \sum_{z \in V_{K+}} \sup_{\theta \in \Theta_{K+}} f(z, A; \theta).
\]
Hence the claim follows.

**Proof of Theorem 2.7.** First note
\[
L_{K,K^+} = \log \sup_{\theta \in \Theta_{K+}} g(A; \theta) f(Z, A; \theta^*) + O_P(1),
\]
where
\[
\log \sup_{\theta \in \Theta_{K+}} g(A; \theta) f(Z, A; \theta^*) \geq \log \sup_{\theta \in \Theta_{K+}} f(Z, A; \theta) f(Z, A; \theta^*) = O_P(1).
\]
Let \( D(\cdot) \) be a diagonal matrix, upper bounding by the maximum,
\[
\log \sup_{\theta \in \Theta_{K+}} g(A; \theta) f(Z, A; \theta^*)
\]
\[
\begin{align*}
& \leq \max_z \sup_{\theta \in \Theta_{K^+}} \log \frac{f(z, A; \theta)}{f(Z, A; \theta^*)} + n \log K^+ \\
& = \max_z \frac{n^2}{2} \left\{ F_1 \left( O(z) / n^2; N(z) / n^2 \right) - F_1 \left( D(p) H^* D(p), pp^T \right) \right\} + O_P(n) \\
& \leq \max_z \frac{n^2}{2} \left| F_1 \left( O(z) / n^2; N(z) / n^2 \right) - F_1 \left( R^* R^T(z), R11^T R^T(z) \right) \right| \\
& \quad + \max_z \frac{n^2}{2} \left[ F_1 \left( R^* R^T(z), R11^T R^T(z) \right) - F_1 \left( D(p) H^* D(p), pp^T \right) \right] + O_P(n) \\
& \leq C\mu_n \max_z \left\| \frac{O(z)}{\mu_n} - RS^* R^T \right\|_\infty + O_P(n) \\
& = O_P(n^{3/2} \rho_1^{1/2}) \\
\end{align*}
\]

using (A.1) in Lemma A.1, and the fact that
\[
\max_{z \in [K^+]^n} F_1 \left( R^* R^T(z), R11^T R^T(z) \right) = F_1 \left( D(p) H^* D(p), pp^T \right).
\]

Next we prove Theorem 2.11.

**Proof of Theorem 2.11.** It remains to upper bound $L_{K, K^+}$. By Lemma 2.6 and Assumption 2.9, it suffices to consider
\[
\max_{z \in V_{K^+}} \sup_{\theta \in \Theta_{K^+}} \log f(z, A; \theta) - \sup_{\theta \in \Theta_{K^+}} \log g(A; \theta) \\
= \max_{z \in V_{K^+}} \sup_{\theta \in \Theta_{K^+}} \log f(z, A; \theta) - \log f(Z, A; \theta^*) + O_P(1).
\]

It follows from the definition of $V_{K^+}$ there exists a surjective function $h : [K^+] \to [K]$ describing the block assignments in $R(z, Z)$. We have
\[
\begin{align*}
\max_{z \in V_{K^+}} \sup_{\theta \in \Theta_{K^+}} & \log f(z, A; \theta) - \log f(Z, A; \theta^*) \\
= & n \sum_{k=1}^{K^+} \alpha \left( n_k(z) / n \right) - n \sum_{a=1}^{K} \alpha \left( \sum_{k \in h^{-1}(a)} n_k(z) / n \right) \\
& + \frac{1}{2} \sum_{k=1}^{K^+} \sum_{l=1}^{K^+} \left( O_{k,l} \log \frac{\hat{H}_{k,l}}{\hat{H}_{h(k),h(l)}} + (n_{k,l} - O_{k,l}) \log \frac{1 - \hat{H}_{k,l}}{1 - \hat{H}_{h(k),h(l)}} \right), \\
\end{align*}
\]

(A.22)

where $\hat{H}_{k,l} = O_{k,l}(z) / n_{k,l}(z)$. The first part of the expression is nonpositive since $\alpha$ is superadditive.
For \( z \in \mathcal{N}_{K^+} \), \( \hat{H}_{k,l} - H^*_h(k),h(l) = O_P(n^{-1} \rho_n^{1/2}) \). Furthermore, the order is uniform since by (A.14),
\[
\| X(z) - X(z_0) \|_{\infty} = o_P(1)
\]
for any fixed \( z_0 \in \mathcal{N}_{K^+} \), and all \( z \in \mathcal{N}_{K^+} \), \( z \notin S(z_0) \). It follows by Taylor expansion that (A.22) is upper bounded by
\[
\frac{1}{4} \sum_{k,l} n_{k,l} \frac{(\hat{H}_{k,l} - H^*_h(k),h(l))^2}{H^*_h(k),h(l)} + o_P(1) = O_P(1)
\]
uniformly for all \( z \in \mathcal{N}_{K^+} \). The claim follows with Assumptions 2.9 and 2.10. Since (A.22) has \( K^+(K^+ + 1)/2 \) terms of order \( O_P(1) \), it suffices to bound the model complexity term for \( \sup_{\theta \in \Theta_{K^+}} \log g(A; \theta) \) by \( \lambda \cdot K^+(K^+ + 1)/2 \cdot \log n \) for some constant \( \lambda \).

\[\square\]

ACKNOWLEDGEMENTS

The authors thank Haiyan Huang for helpful discussions. This research was funded in part by NSF FRG Grant DMS-1160319.

REFERENCES

Airoldi, E. M., Blei, D. M., Fienberg, 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. Stat. 41 2097-2122.
Bickel, P. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. Proc. Natl. Acad. Sci. USA 106 21068-73.
Bickel, P. J. and Sarkar, P. (2013). Hypothesis testing for automated community detection in networks. arXiv preprint arXiv:1311.2694.
Bickel, P., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. Ann. Stat. 41 1922-1943.
Cérou, A., Daudin, J.-J., Pierre, L. et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electron J. Stat. 6 1847–1899.
Chen, K. and Lei, J. (2014). Network cross-validation for determining the number of communities in network data. arXiv preprint arXiv:1411.1715.
Choi, D. S., Wolfe, P. J. and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. Biometrika asr053.
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 Zdeborová, 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 partitioning for the stochastic block model when the model parameters are unknown. SIAM J. Matrix Anal. Appl. 34 23–39.
Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. Soc. Networks 5 109-137.
Joseph, A. and Yu, B. (2013). Impact of regularization on Spectral Clustering. arXiv preprint arXiv:1312.1733.
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.
Lei, J. (2014). A Goodness-of-fit Test for Stochastic Block Models. arXiv preprint arXiv:1412.4857.
Leskovec, J. and McAuley, J. J. (2012). Learning to discover social circles in ego networks. In Adv. Neural Inf. Process. Syst. 539–547.
Newman, M. E. J. (2006a). Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA 103 8577–8582.
Newman, M. E. (2006b). Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74 036104.
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. Stat. 39 1878-1915.
Saldana, D. F., Yu, Y. and Feng, Y. (2014). How Many Communities Are There? arXiv preprint arXiv:1412.1684.
Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. arXiv preprint arXiv:1309.5936.
Zhao, Y., Levina, E. and Zhu, J. (2011). Community extraction for social networks. Proc. Natl. Acad. Sci. USA 108 7321–7326.

Y.X.R. Wang and P.J. Bickel
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley, California, 94720
E-mail: ryxwang@gmail.com
bickel@stat.berkeley.edu