Network Cross-Validation for Determining the Number of Communities in Network Data

Kehui Chen[1] and Jing Lei[2]

[1]Department of Statistics, University of Pittsburgh
[2]Department of Statistics, Carnegie Mellon University

November 7, 2014

Abstract

The stochastic block model and its variants have been a popular tool in analyzing large network data with community structures. Model selection for these network models, such as determining the number of communities, has been a challenging statistical inference task. In this paper we develop an efficient cross-validation approach to determine the number of communities, as well as to choose between the regular stochastic block model and the degree corrected block model. Our method, called *network cross-validation*, is based on a block-wise edge splitting technique, combined with an integrated step of community recovery using sub-blocks of the adjacency matrix. The solid performance of our method is supported by theoretical analysis of the sub-block parameter estimation, and is demonstrated in extensive simulations and a data example. Extensions to more general network models are also discussed.

1 Introduction

In the last few decades, the amount of network data and the need for relevant statistical inference tools are growing at a rapid pace. In network data analysis, the observed data is a graph over n nodes, where each node represents an individual in a population, and an edge between two nodes represents the presence of a certain kind of relationship or interaction between the individuals. One of the main research topics in network data analysis is to identify hidden communities from a single observed network. Roughly speaking, network community refers to the phenomenon that individuals close to each other are more likely to connect, and hence the edge density varies from within coherent subpopulations to between subpopulations (Newman & Girvan, 2004; Newman, 2006). The stochastic block model (Holland et al., 1983) and its variants such as the degree corrected block model (Karrer & Newman, 2011) are powerful and mathematically elegant tools to model large networks

with community structures, and have been proved useful in many scientific areas such as social science, biology, and information science (Faust & Wasserman, 1992; Kemp et al., 2006; Bickel & Chen, 2009).

The community recovery problem for stochastic block models has been the focus of much research effort in the past decade, in several areas including statistics (Bickel & Chen, 2009; Zhao et al., 2012; Jin, 2012; Fishkind et al., 2013; Lei & Rinaldo, 2013), machine learning (McSherry, 2001; Chen et al., 2012; Chaudhuri et al., 2012; Anandkumar et al., 2014), statistical physics (Decelle et al., 2011; Krzakala et al., 2013), and probability theory (Massoulie, 2013; Mossel et al., 2013; Abbe et al., 2014). These methods are based on a wide range of different tools such as maximum likelihood, convex optimization, spectral methods, and belief propagation, etc. However, almost all of these methods require K, the total number of communities, to be known in advance.

Unlike the community recovery problem, there has been much less development on determining the number of communities. For stochastic block models, the null model corresponding to $K = 1$ is an Erdős-Rényi graph. Zhao et al. (2011) propose to sequentially extract one significant community from the remaining of the network, and they approximate the null distribution of their optimizing statistic by bootstrapping from an Erdős-Rényi graph. Bickel & Sarkar (2013) propose to test $K = 1$ vs $K > 1$ at each step of a recursive bipartition algorithm. They derive the asymptotic null distribution of the largest eigenvalue of the suitably scaled and centered adjacency matrix. But the convergence rate is slow and an empirical tuning is needed in practice. Also it requires a diagonal dominant condition when examining the power. Moreover, these sequential or recursive testing procedures only work for certain types of community structures. To directly test $K = k$ vs $K > k$ remains an open problem due to the difficulty of approximating the null distribution.

In standard model-based clustering, data points are assumed to be independently drawn from a common underlying mixture distribution, and it is often possible to determine the number of mixture components by modifying the AIC- and BIC-type model selection criteria. For stochastic block models, it is hard to determine the model complexity using suitable quadratic approximations of the log likelihood at the empirical maximum, which is a key step in deriving AIC and BIC. Another popular method in general model selection problems is cross-validation. In the traditional formulation of stochastic block models, each node i ($1 \le i \le n$) is assigned a community membership $g_i \in \{1, ..., K\}$ independently with probability $P(g_i = k) = \pi_k$ for $1 \le i \le n$, $1 \le k \le K$. Therefore, a stochastic block model is parameterized by the membership distribution $\pi = (\pi_1, ..., \pi_K)$ and the community-wise edge probability matrix $B \in [0, 1]^{K \times K}$. Following the traditional parameterization, a straightforward cross-validation method for the stochastic block model has been considered before, where one estimates π and B on the adjacency matrix confined on a training subset of nodes, and evaluates the fitting on the adjacency matrix confined on the testing subset of nodes. This approach is rarely used in practice, due to several drawbacks. First,

calculating the full likelihood in presence of a missing membership vector g is computationally demanding. Second, it introduces unnecessary randomness in the validation step by treating the node memberships as random variables. Third, it does not use the observed edges between the training and testing nodes, which contain useful information for inference. As we will see below, all these problems can be resolved by a novel *network cross-validation* (NCV) approach proposed in this paper.

The NCV method developed in this paper is based on a block-wise edge splitting technique, combined with the idea of integrating community recovery into model selection. Given $n_1 < n$, consider a block-splitting of the adjacency matrix

$$A = \begin{pmatrix} A^{(11)} & A^{(12)} \\ A^{(21)} & A^{(22)} \end{pmatrix}, \tag{1}$$

where $A^{(11)}$ is an $n_1 \times n_1$ principal submatrix chosen at random. Our method first estimates the community-wise edge probability B and a full membership vector $g = (g_i : 1 \le i \le n)$ from the $n_1 \times n$ rectangular matrix $A^{(1)} = (A^{(11)}, A^{(12)})$. Many standard procedures designed for the full adjacency matrix can be extended to this case, such as likelihood based methods and spectral methods. Our empirical and theoretical results are based on spectral clustering. After obtaining the estimates $(\hat{g}, \hat{B})$ from the rectangular matrix, we can assess the goodness-of-fit of the estimated parameter by validating on the testing set of edges contained in $A^{(22)}$. A natural choice of the predictive loss function ℓ is the negative log-likelihood $\ell(x, p) = -x \log p - (1 - x) \log(1 - p)$.

A key innovation of the proposed block-wise edge splitting techniques is that we train the model on a rectangular submatrix that carries the full structural information of the network. Therefore, the memberships for n nodes and the community-wise edge probability matrix B can be consistently estimated from the training set (Theorem 1 in Section 2). This distinguishes the proposed NCV method from the traditional cross-validation perspective where the splitting is on nodes. The proposed NCV approach takes advantage of the conditional independence between the two subsets of edges given the community partition in a stochastic block model. It reflects a significant difference between the traditional parameterization of the stochastic block model that treats the node memberships as random variables, and the conditional parameterization that treats the memberships as parameters. The traditional parameterization can model networks of arbitrary size and has a motivation from exchangeable random graphs (Bickel & Chen, 2009). However, given an observed realization of the stochastic block model, the useful information for statistical inference is largely contained in the randomness of edge formulation, and it is usually beneficial to use conditional inference by treating the memberships as fixed parameters. In particular, when community recovery is of interest, these two parameterizations do not exhibit much difference, but it is more natural to consider the memberships as fixed parameters to be estimated. For example, the conditional parameterization is used in the profile likelihood

method proposed in Bickel & Chen (2009), and similarly in spectral clustering (Lei & Rinaldo, 2013).

The NCV method can be applied to select the best model from a general collection of candidate models, which do not need to be nested or hierarchical. For example, one can use NCV to choose between the regular stochastic block model and the degree corrected block model, with simultaneous choice of number of communities. Moreover, the block-wise edge splitting idea behind NCV can be further generalized to other network models with conditional edge independence. These extensions are described in Section 3 and Section 5, and illustrated in an application to a political blog data in Section 4, where the NCV method chooses the degree corrected block model with two communities, matching pervious findings in the literature.

So far there is no widely accepted approach in choosing the number of communities in the network literature, and we believe the proposed NCV method is of substantial practical interest. In Section 4, we demonstrate the effectiveness of our method via extensive simulations, where different types of network community structures are investigated. The proposed NCV method is tuning free except the number of folds. Throughout the paper, we use three-fold NCV, which is computationally efficient and has very satisfactory empirical performance under moderate sample sizes. Further discussions can be found in Section 5.

2 Network cross-validation for stochastic block models

In a stochastic block model with n nodes and K communities, the observed random graph is often represented by a n by n symmetric binary adjacency matrix A. The community structure is represented by a vector $g \in \{1, ..., K\}^n$ with g_i being the community that node i belongs to. Given the membership vector g, each edge A_{ij} $(i < j)$ is an independent Bernoulli variable satisfying

$$P(A_{ij} = 1) = 1 - P(A_{ij} = 0) = B_{g_i g_j}\,, \tag{2}$$

where $B \in [0, 1]^{K \times K}$ is a symmetric matrix representing the community-wise edge probability. In this section we focus on the problem of estimating K, the number of communities, from a single observed network A. Generalization to other model selection problems is straightforward and will be discussed in later sections.

2.1 Block-wise edge splitting

Given the community membership g, the only randomness in the observed graph is the edge formulation. Our NCV approach is based on block-wise edge splitting. Let $(\mathcal{N}_1, \mathcal{N}_2)$

be a partition of the nodes, the adjacency matrix can be written in a block form as in (1), where $A^{(jj)}$ is the adjacency matrix for nodes in $\mathcal{N}_j$ $(j = 1, 2)$.

The splitting step puts edges in $A^{(11)}$ and $A^{(12)}$ in the training sample and $A^{(22)}$ the testing sample. Such a block-wise edge splitting is neither node splitting nor simple edge splitting. It originates from two key observations. First, the rectangular matrix $A^{(1)} = (A^{(11)}, A^{(12)})$ carries information on the community structure of the entire network. That is, we can estimate the membership of each of the n nodes as well as the community-wise edge probability matrix from $A^{(1)}$. Second, given the community membership g, these two sets of edges in $A^{(1)}$ and $A^{(22)}$ are independent. Generalization of this block-wise edge splitting procedure to models other than stochastic block models is feasible, as long as these two key facts hold.

Such a block-wise edge splitting makes full use of the entire observed adjacency matrix and provides a way to directly compare multiple values of K based on the predictive loss on the testing sample. We summarize the key component of the NCV procedure as follows.

Algorithm 1: model evaluation using sample splitting

Input: adjacency matrix A, number of communities K, training block size n_1.

1. Randomly split the adjacency matrix into $(A^{(11)}, A^{(12)}; A^{(12)}, A^{(22)})$ as in (1), where $A^{(11)}$ contains edges between nodes in a random subset $\mathcal{N}_1$ of size n_1, $A^{(22)}$ contains edges between nodes in $\mathcal{N}_2$, the compliment of $\mathcal{N}_1$, and $A^{(12)}$ contains edges between $\mathcal{N}_1$ and $\mathcal{N}_2$.
2. Estimate model parameters $(\hat{g}, \hat{B})$ using the rectangular submatrix $A^{(1)} = (A^{(11)}, A^{(12)})$.
3. Output the predictive loss evaluated on $A^{(22)}$:

$$\hat{L}(A, K) = \sum_{i,j \in \mathcal{N}_2, i \neq j} \ell(A_{ij}, \hat{P}_{ij}) \,,$$

where ℓ is a loss function and $\hat{P}_{ij}$ is an estimate of $P_{ij} = E(A_{ij})$.

We give further details on how to adapt this sample splitting validation method to a V-fold network cross-validation in Section 2.4 below. In the following we discuss Steps 2 and 3 of Algorithm 1 in further detail.

2.2 Estimating model parameters from the rectangular matrix

Algorithm 1 estimates model parameters (g, B) from the $n_1 \times n$ rectangular matrix $A^{(1)} = (A^{(11)}, A^{(12)})$. Many standard procedures designed for the full adjacency matrix can be extended to this case, such as likelihood based methods and spectral methods. Here we focus on spectral clustering, because it is simple to implement and the analysis is

straightforward. This is also the method we implement in the numerical experiments presented in Section 4.

The simple spectral clustering method first performs a singular value decomposition on $A^{(1)}$, and estimates g by applying k-means clustering on the rows of the $n \times K$ matrix consisting of the leading K right singular vectors. Once $\hat{g}$ is obtained, let $\mathcal{N}_{j,k}$ be the nodes in $\mathcal{N}_j$ with estimated membership k, and $n_{j,k} = |\mathcal{N}_{j,k}|$ $(j = 1, 2,\ 1 \le k \le K)$. We can estimate B using a simple plug-in estimator:

$$\hat{B}_{k,k'} = \begin{cases} \dfrac{\sum_{i \in \mathcal{N}_{1,k}, j \in \mathcal{N}_{1,k'} \cup \mathcal{N}_{2,k'}} A_{ij}}{n_{1,k}(n_{1,k'} + n_{2,k'})}, & k \neq k', \\ \dfrac{\sum_{i,j \in \mathcal{N}_{1,k}, i<j} A_{ij} + \sum_{i \in \mathcal{N}_{1,k}, j \in \mathcal{N}_{2,k}} A_{ij}}{(n_{1,k} - 1)n_{1,k}/2 + n_{1,k}n_{2,k}}, & k = k'. \end{cases} \tag{3}$$

When K is the true number of blocks, Theorem 1 provides performance guarantee for $(\hat{g}, \hat{B})$, in an asymptotic sense when n increase to ∞, under the following three conditions.

(A1) $B = \alpha_n B_0$, where α_n is a scaling factor, possibly changing with n, and B_0 is a $K \times K$ fixed symmetric matrix with full rank.

(A2) The smallest community size is at least $\pi_0 n$ for a constant $\pi_0 \in (0, 1)$.

(A3) The training block size $n_1 \ge c_0 n$ for some constant c_0.

Theorem 1 (Consistency of parameter estimation)**.** Under assumptions (A1–A3), $(\hat{g}, \hat{B})$, the parameters estimated from $A^{(1)}$ using spectral clustering followed by plug-in as described above, satisfies
(a) if $\alpha_n \ge c \log n/n$ for a large enough constant c, then with probability tending to one $\hat{g}$ agrees with g on all but $O(\alpha_n^{-1})$ nodes;
(b) if $\alpha_n^{-1} = o(n^{1/2})$, then $\hat{B} = B(1 + o_P(1))$.

Remark 1. The proof of Theorem 1 is given in Appendix A. Part (a) establishes the consistency of community recovery when the expected node degrees is of order $\log n$ or higher. Lei & Rinaldo (2013) have developed the same result for simple spectral clustering applied to the full adjacency matrix. Our consistency result does not need to assume that the k-means step achieves the global maximum. As proved in Lei & Rinaldo (2013), polynomial-time approximate k-means algorithms can be used and lead to essentially the same theoretical guarantee, and the proof can be adapted to our case. The consistency for B matrix requires the expected node degrees to grow faster than $\sqrt{n}$ order. This condition can be slightly weakened if other methods are used to estimate g, such as those developed in Vu (2014) and Bickel & Chen (2009). We use simple spectral clustering throughout this paper, but the proposed NCV method is based on a generic block-wise sample splitting idea, which can be combined with many other community recovery algorithms.

2.3 Validation using the testing set of edges

After estimating the parameters $(\hat{g}, \hat{B})$ in Step 2, we can assess the goodness-of-fit by validating on the testing set of edges.

For each edge in the testing set, A_{ij} $(i \neq j,\, i, j \in \mathcal{N}_2)$ is a Bernoulli random variable with parameter $P_{ij} = B_{g_i g_j}$, which is approximated by $\hat{P}_{ij} = \hat{B}_{\hat{g}_i \hat{g}_j}$. Some natural choices of the loss function ℓ in Step 3 include negative log-likelihood $\ell(x, p) = -x \log p - (1-x)\log(1-p)$, and squared error $\ell(x, p) = (x-p)^2$. In our numerical experiments, these two loss functions give almost identical performance, so we will focus on the log-likelihood loss function.

In the validation step, if K is too small, then the fitted model cannot capture the fine structures in the data, and will likely lead to poor predictive loss on testing data. If K is too large, then the model overfits the training data, with noisy prediction on the testing data. Therefore, it is natural to expect the validated predictive loss $\hat{L}(A, K)$ to be minimized when K is the true number of communities.

2.4 V-fold network cross-validation

Now we formally describe the V-fold network cross-validation procedure.

Algorithm 2: V-fold network cross-validation

Input: adjacency matrix A, a set $\mathcal{K}$ of candidate values for K, number of folds $V \geq 2$.

1. Randomly split the adjacency matrix into $V \times V$ equal sized blocks

$$A = (A^{(ij)} : 1 \leq i, j \leq V)$$

similarly as in (1), where the nodes are partitioned into V equal-sized subsets $\mathcal{N}_j$ $(1 \leq j \leq V)$; $A^{(jj)}$ contains edges between nodes in the jth random subset $\mathcal{N}_j$; and $A^{(ij)}$ contains edges between $\mathcal{N}_i$ and $\mathcal{N}_j$.

2. For each $1 \leq j \leq V$, and each $K \in \mathcal{K}$

   (a) Estimate model parameters $(\hat{g}^{(j)}, \hat{B}^{(j)})$ using the rectangular submatrix obtained by removing the rows of A in subset $\mathcal{N}_j$

$$A^{(-j)} = (A^{(il)} : i \neq j,\ 1 \leq i, l \leq V)\,.$$

   (b) Calculate the predictive loss evaluated on $A^{(jj)}$:

$$\hat{L}^{(j)}(A, K) = \sum_{u,v \in \mathcal{N}_j, u \neq v} \ell(A_{uv}, \hat{P}^{(j)}_{uv})\,,$$

   where $\hat{P}^{(j)}_{uv} = \hat{B}^{(j)}_{\hat{g}^{(j)}_u, \hat{g}^{(j)}_v}$.

3. Let $\hat{L}(A,K) = \sum_{j=1}^{V} \hat{L}^{(j)}(A,K)$ and output

$$\hat{K} = \arg\min_{K \in \mathcal{K}} \hat{L}(A,K)\,.$$

In our experiments we found the performance of NCV insensitive to the choice of V, and we used $V = 3$ for all numerical experiments. Further discussion on the choice of V and its difference from the regular cross-validation is given in Section 5.

3 Degree corrected block models and further extensions

3.1 Choosing K for degree corrected block models

The degree corrected block model (Karrer & Newman, 2011) is a generalization of the stochastic block model. Given membership vector g and community-wise connectivity matrix B, the presence of an edge between nodes i and j is represented by a Bernoulli random variable A_{ij} with

$$P(A_{ij} = 1) = 1 - P(A_{ij} = 0) = \psi_i \psi_j B_{g_i g_j}, \tag{4}$$

where $\psi_i > 0$ represents the *individual activeness* of node i. Thus the degree corrected block model is parameterized by a triplet (g, B, ψ), with identifiability constraint $\max_{i:g_i=k} \psi_i = 1$ for all $k = 1, ..., K$. The regular stochastic block model is a special case with $\psi_i = 1$ for all i. Recently, efficient community recovery methods have been developed for degree corrected block models with high accuracy under mild conditions (see, for example, Zhao et al., 2012; Jin, 2012; Chaudhuri et al., 2012; Lei & Rinaldo, 2013). We now extend the procedure described in Section 2 to degree corrected block models.

The framework given in Algorithm 1 is general enough to cover the degree corrected block model. The implementation needs to be changed in Step 2 because regular spectral clustering may not work well for degree corrected block models, and we also need to estimate the node activeness parameter ψ. To this end, we consider a spherical spectral clustering method.

Spherical spectral clustering:

Input: Rectangular $n_1 \times n$ matrix $A^{(1)}$, number of communities K.

1. Let $\hat{U}$ be the $n \times K$ matrix consisting of the top K right singular vectors of $A^{(1)}$.
2. Let $\tilde{U}$ be the matrix obtained by scaling each row of $\hat{U}$ to unit norm.
3. Output $\hat{g}$ by applying the k-median clustering algorithm to the rows of $\tilde{U}$.

The normalization step in the spherical spectral clustering algorithm decouples the effect of node activeness ψ from the community structure. As shown in the proof of Theorem 2 below, the community information is contained in the normalized matrix $\tilde{U}$, whereas the node activeness information is contained in the row norms of $\hat{U}$.

The community recovery is obtained by a k-median clustering algorithm, which finds a collection of center points to minimize the sum of ℓ_2 distance from each data point to its nearest center, instead of the squared ℓ_2 distance as in the k-means. To be precise, given input matrix $\tilde{U}$ and number of centers K, the k-median clustering solves, possibly with approximation, the following optimization problem:

$$\min_{v_1,\ldots,v_K\in\mathbb{R}^K,\ g\in\{1,\ldots,K\}^n} \sum_{i=1}^{n} \|\tilde{u}_i - v_{g_i}\|\,,$$

where $\tilde{u}_i$ is the ith row of $\tilde{U}$. Approximate solutions within a constant factor from the global optimum can be found using efficient algorithms (Charikar et al., 1999; Li & Svensson, 2013). Our theoretical analysis is applicable to such approximate solutions. If the matrix $\hat{U}$ has zero rows, one can apply the spherical clustering algorithm on the non-zero rows and assign arbitrary membership to the zero rows. Our theoretical analysis shows that with high probability the number of zero rows in $\hat{U}$ is negligible under mild conditions.

To estimate the node activeness parameter ψ, let

$$\hat{\psi}'_i = \ell_2 \text{ norm of the } i\text{th row of } \hat{U}\,,$$

and $\psi' = (\psi'_i : 1 \le i \le n)$ with

$$\psi'_i = \frac{\psi_i}{\sqrt{\sum_{j:g_j=g_i} \psi_j^2}}$$

be the community-normalized version of ψ. We will show, in the proof of Theorem 2 below, that $\hat{\psi}'$ is a good estimate of ψ' under appropriate conditions. Due to the scaling identifiability of ψ and B, having a good estimate of ψ' is sufficient for our purpose and one can proceed with the plug-in estimator:

$$\hat{B}'_{k,k'} = \begin{cases} \dfrac{\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} A_{ij}}{\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} \hat{\psi}'_i\hat{\psi}'_j}\,, & k \ne k\,, \\[2ex] \dfrac{\sum_{i,j\in\mathcal{N}_{1,k}, i<j} A_{ij} + \sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{2,k}} A_{ij}}{\sum_{i,j\in\mathcal{N}_{1,k}, i<j} \hat{\psi}'_i\hat{\psi}'_j + \sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{2,k}} \hat{\psi}'_i\hat{\psi}'_j}\,, & k = k'\,. \end{cases}$$

The estimated $P_{ij} = E(A_{ij})$ to be used for validation is then

$$\hat{P}_{ij} = \hat{\psi}'_i\hat{\psi}'_j\hat{B}'_{\hat{g}_i,\hat{g}_j}\,.$$

To investigate theoretical properties of these estimators, we assume that there are no overly inactive nodes.

(A4) $\inf_{1\le i\le n}\psi_i \ge \psi_0$ for a positive constant ψ_0.

Then we have the following result analogous to Theorem 1, which is also proved in Appendix A.

Theorem 2. Under (A1)–(A4), we have
(a) if $\alpha_n \ge c\log n/n$ for a large enough constant c, then with probability tending to one $\hat{g}$ agrees with g on all but $O(\sqrt{n/\alpha_n})$ nodes;
(b) if $\alpha_n^{-1} = o(n^{1/3})$ then $\hat{P}_{ij} = P_{ij}(1+o_p(1))$ for all but a vanishing proportion of node pairs.

3.2 Choosing model types and K simultaneously

The above extension to degree corrected block models allows us to compare and choose, for a given adjacency matrix, between the regular stochastic block model and the degree corrected block model. Sometimes it is desirable to tell if the degree heterogeneity in an observed network can be explained by pure random fluctuation in a stochastic block model (see, for example, Yan et al., 2014).

Our V-fold NCV can be used simultaneously to choose between the regular stochastic block model and the degree corrected block model, and to determine the number of blocks. To this end, one just needs to calculate the regular stochastic block model validation error $\hat{L}_{\mathrm{sbm}}(A,K)$, and the degree corrected block model validation error $\hat{L}_{\mathrm{dcbm}}(A,K)$, for a collection of values of K as described in Section 2.4. The best model is chosen by finding the overall smallest cross-validation loss. We illustrate this method on simulated data and on a political blog data in Section 4.

4 Numerical Experiments

In this section, we illustrate the performance of our proposed NCV method by three simulations and one data example.

Simulation 1: edge sparsity and community imbalance. This simulation is designed to investigate the performance of choosing K for stochastic block models under different levels of edge sparsity and community size imbalance. We use the community-wise edge probability matrix $B = rB_0$, where the diagonal entries of B_0 are 3 and off-diagonal entries are 1. The sparsity level is controlled by $r \in (0, 1/3)$. We use a sequence of $r \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, so that for $n = 1000$ the smallest expected degree ranges from

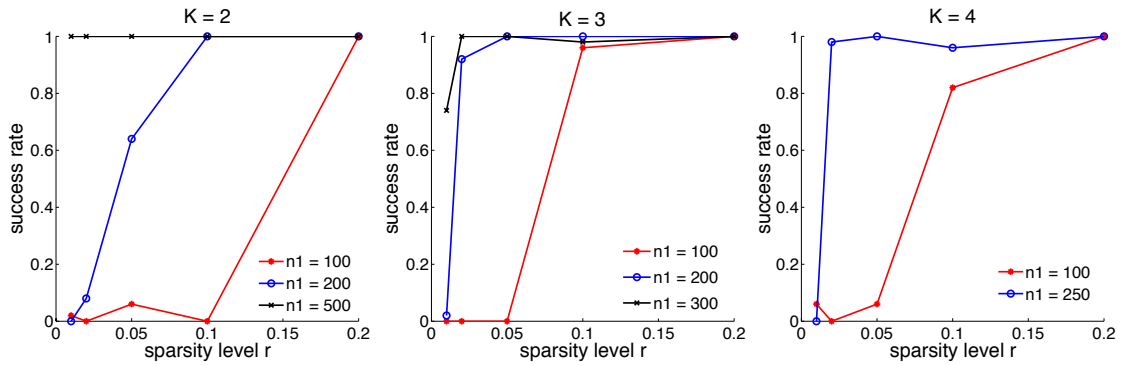

Figure 1: Results for **Simulation 1**: reporting the proportion of correct estimate of K for stochastic block models, for $K = 2, 3, 4$, under various sparsity levels $r \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, and various sizes of the first community n_1. The number of nodes is 1000.

12 to 400. Let n_1 be the size of the smallest community, and the size of each of the remaining $K-1$ communities be $(n-n_1)/(K-1)$. We generate edges according to the stochastic block model (2). For each combination of (r, K, n_1), three-fold NCV model selection is carried out for 50 independently drawn adjacency matrices. Figure 1 shows the proportion of correct model selection among these 50 repetitions as functions of r for different n_1 and $K = 2, 3, 4$. As expected, the performance is better as r and n_1 increase. In particular, for $K = 2$, in the most balanced case where $n_1 = 500$, the proposed NCV can perfectly choose the true number of clusters even for the sparsest case where $r = 0.01$, whereas in the most imbalanced case where $n_1 = 100$, there is a phase transition near $r = 0.1$. The curve for $n_1 = 200$ is in between. The same phenomenon is observed for $K = 3$ and $K = 4$. The proposed NCV can almost perfectly pick out K for relatively balanced community sizes, even for very sparse cases. For imbalanced cases, one needs to have moderate expected degrees for the nodes in the smallest community. We note that community recovery for a given K is an integrated step in the proposed NCV method, so it is expected that the performance of NCV is closely related to the difficulty of the community recovery problem when knowing the true K, which may depend on the particular community recovery method used in NCV.

Simulation 2: general block structures and comparison to recursive bipartition. This simulation is designed to further investigate the proposed NCV method under general block structures of networks, and meanwhile to compare the proposed NCV method with the recursive testing procedure proposed in Bickel & Sarkar (2013). We generate symmetric B randomly as follows. For each upper-triangle entry of B, we generate a random number from $\text{Unif}(0, 0.5)$. The upper bound 0.5 is set to exclude unrealistically dense networks

that are of less interest. We only use B matrices whose Kth singular values are in the upper three quarters and therefore have relatively well-formed K-block structures. The membership vector g is generated from multinomial distribution (n, π) with equal probability $\pi = (1/K, \ldots, 1/K)$. For each simulated data, we applied three-fold NCV method as well as the recursive bipartition algorithm developed in Bickel & Sarkar (2013) with $\alpha = 0.01$. The basic idea of the recursive bipartition method is to divide the nodes into two clusters if $K = 1$ is rejected at level α, and then recursively test $K = 1$ vs $K > 1$ on each of the two sub-networks until failing to reject $K = 1$. The success rates in 50 simulations for each combination of $n = 600, 1200$ and $K = 1, 2, 3, 4$ are shown in Figure 2. As expected, both methods benefit from a larger sample size (top row vs bottom row). The task of determining K gets harder as the true number of communities gets larger (from left column to right column). The proposed NCV method performs uniformly better than the bipartition method. The NCV method is also much faster than the bipartition method, where the latter requires a small bootstrap sample to adjust the null distribution at each testing step. The simulation design suggests that the proposed NCV has very satisfactory performance under very general structures of B. For $n = 1200$, the empirical success rate of NCV achieves 100% for $K = 1,\ 2$, 84% for $K = 3$, and 72% for $K = 4$.

Simulation 3: degree corrected block models. This simulation is designed to demonstrate the performance of selecting between the stochastic block model and the degree-corrected block model with simultaneous selection of K. We use a B matrix whose diagonal is 0.25 and off-diagonal is 0.1, which gives a moderate sparsity level for stochastic block models. For degree-corrected block model, the degree parameter ψ is generated from $\mathrm{Unif}(0.2, 1)$, and normalized to have block-wise maximum value 1. The edges are generated according to (4). The network is much sparser in presence of the degree parameter ψ and the inference problem is harder. Three-fold NCV is used to simultaneously choose the model type T from $T =$ "SBM" or $T =$ "DCBM", and the number of communities K. Table 1 shows the proportion of correct model type selection $\hat{T} = T$ and proportion of correct choice of K given correct model type selection. Data are generated 50 times from both the stochastic block model and the degree corrected block model, for each combination of $K = 1, 2, 3, 4$ and $n = 300, 600, 1200$. We observe that when the true model type is stochastic block model, NCV can almost perfectly pick out the correct model and correct K for various combinations of K and n. As expected, a relatively larger sample size is needed to get good performance when the network is generated from a degree corrected block model. Our simulation shows that for $n = 1200$, NCV can almost always pick out the correct DCBM model with the right K.

Data example: political weblogs. The political blog data was collected and analyzed in Adamic & Glance (2005). The data set contains snapshots of over one thousand weblogs shortly before the 2004 U.S. Presidential Election, where the nodes are weblogs, and edges are hyperlinks. The nodes are labeled as being either liberal or conservative, which can be treated as two well-defined communities. The degree corrected block model is

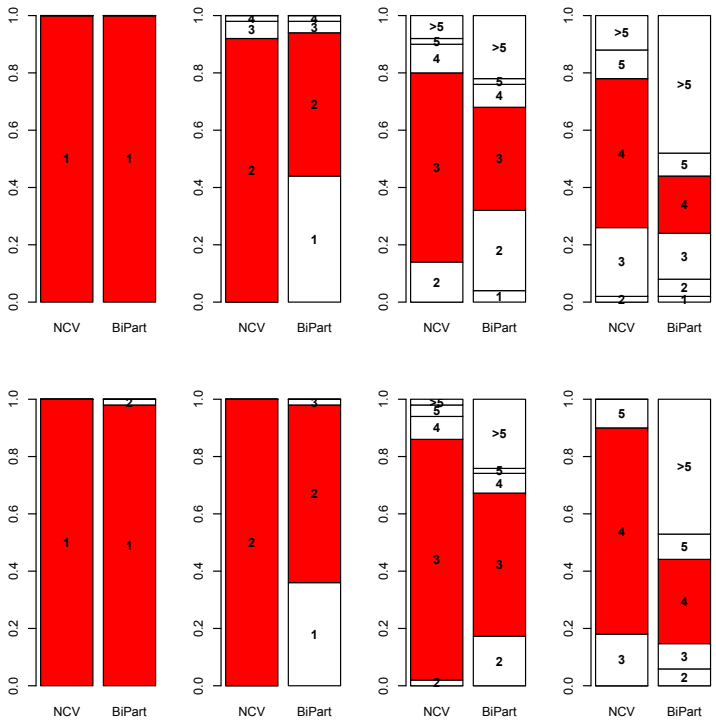

Figure 2: Results for **Simulation 2**: reporting the proportion of selected K by NCV (the proposed method) and BiPart (Bickel & Sarkar (2013)), for true $K = 1, 2, 3, 4$ (from left to right), and sample size $n = 600$ (top row) and $n = 1200$ (bottom row).

believed to fit better than the stochastic block model to this data with two communities (Karrer & Newman, 2011; Zhao et al., 2012; Jin, 2012). To illustrate the NCV method for simultaneously choosing between the regular stochastic block model and the degree corrected block model, and choosing the number of communities K, we apply three-fold NCV to the largest connected component in the network which contains 1222 nodes. The NCV method consistently selects the degree corrected block model with two communities. The cross-validated negative log-likelihood for all candidate models is plotted in Figure 3 for a typical block splitting. We repeated the NCV selection 100 times using independent random block splittings. The NCV method selected DCBM and $K = 2$ in 99 out of 100 repetitions, where the one failure was due to non-convergence of k-means in spectral clustering.

Table 1: Results for **Simulation 3**: proportion of selecting the correct model type, and choosing the correct K given correct model type selection, from 50 independent simulations. The true models are generated from stochastic block models (SBM) or degree corrected block models (DCBM), for true $K = 1, 2, 3, 4$ and $n = 300, 600, 1200$.

		SBM				*DCBM*			
		$K=1$	2	3	4	$K=1$	2	3	4
$n=300$	$\hat{P}(\hat{T}=T)$	1	1	1	1	1	0.68	0.44	0.42
	$\hat{P}(\hat{K}=K\|\hat{T}=T)$	1	1	0.98	0.92	1	0.41	0	0
$n=600$	$\hat{P}(\hat{T}=T)$	1	1	1	1	1	1	0.96	0.98
	$\hat{P}(\hat{K}=K\|\hat{T}=T)$	1	1	1	0.98	1	1	0.42	0
$n=1200$	$\hat{P}(\hat{T}=T)$	1	1	1	1	1	1	1	1
	$\hat{P}(\hat{K}=K\|\hat{T}=T)$	1	1	1	0.98	1	1	1	1

5 Discussion

Further extensions In general, the network cross-validation approach proposed in this paper is applicable to network models where (i) edges form independently given an appropriate set of model parameters; and (ii) the edge probabilities can be estimated accurately using a subset of rows of the adjacency matrix. The stochastic block model and the degree corrected model are good examples satisfying these two properties. There are other popular network models in this category, such as the random dot-product graph. The random dot-product graph model (Young & Scheinerman, 2007) assumes that each node i has an embedding v_i on a subset of the d-dimensional unit sphere, and that given the embedding the edge between node i and node j is an independent Bernoulli random variable with parameter $\langle v_i, v_j \rangle$. This is a special case of the latent space model (Hoff et al., 2002). The latent vectors can be accurately estimated using spectral methods (Sussman et al., 2013), which can be adapted naturally so that the model parameters can be estimated using only the training subset of rows of A.

Effect of the number of folds In general, cross validation methods are insensitive to the number of folds. The same intuition empirically holds true for the proposed NCV method. However, there is slight difference between the NCV framework and the traditional cross-validation. Unlike traditional cross validation where each data point is included in a testing sample, In V-fold NCV only the diagonal blocks are used as testing samples and hence the proportion of edges included in testing samples is roughly $1/V$. On the other hand, the ratio between the sizes of training and testing samples in a single fold is $(V^2 - 1)$ to 1 for NCV, and $(V - 1)$ to 1 for traditional cross-validation. Roughly speaking, having a larger value of V will rapidly increase the estimation accuracy in the training stage but will

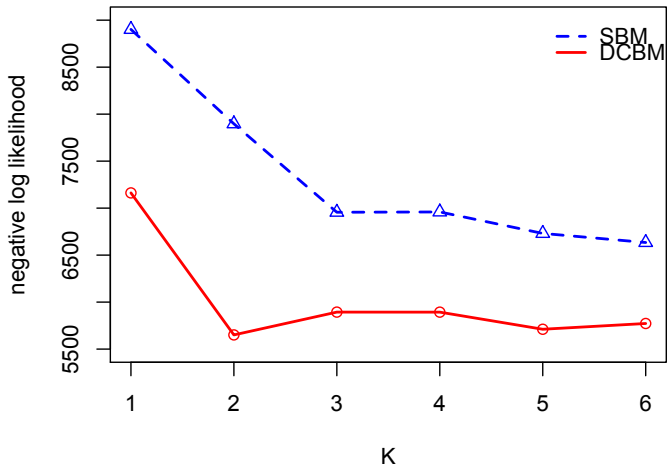

Figure 3: Results for **the political blogs data**: reporting the three-fold cross-validated negative log-likelihood of all candidate models from one random block splitting. Dashed line: stochastic block models; solid line: degree corrected block models. The results are consistent over 100 repeated random block splittings.

reduce the testing sample size. In our numerical experiments we found $V = 3$ a reasonable choice for most cases, which is roughly comparable to a 9-fold traditional cross validation in terms of the training and testing sample size ratio.

A Proofs

Let P be the $n \times n$ matrix such that $P_{ij} = B_{g_i g_j}$, and $P^{(1)}$ be defined using the same block representation of A. Let $\mathcal{N}^*_{j,k}$ be the nodes in subsample $\mathcal{N}_j$ belonging to community k and $n^*_{j,k} = |\mathcal{N}^*_{j,k}|$ $(j = 1, 2,\ k = 1, ..., K)$. For any matrix M, let $\sigma_K(M)$ be its Kth largest singular value. In the statement of results and the proof, constants c, C may take different values from line to line. We let $\|M\| = \sigma_1(M)$ be the spectral norm of M and $\|M\|_F = (\sigma_1^2(M) + \sigma_2^2(M) + ...)^{1/2}$ be the Frobenious norm.

Lemma 3 (Size of split community)**.** Under Assumptions (A2) and (A3), for n large enough we have $\min_k n^*_{1,k} \geq c_0\pi_0 n/2$, with probability at least $1 - n^{-1}/2$.

The proof of this lemma follows from a simple application of large deviation bounds for hypergeometric random variables (Skala, 2013) combined with union bound and is omitted.

Lemma 4 (Spectral norm error of partial adjacency matrix)**.** Let A be the adjacency matrix generated from a degree corrected block model satisfying Assumptions A1 and A4, with $\alpha_n \geq c \log n/n$ for a positive constant c. Let $A^{(1)}$ be an arbitrary subset of rows of A and $P^{(1)}$ be the corresponding submatrix of P. We have, for some constant C,

$$P\left(\|A^{(1)} - P^{(1)}\| \leq C\sqrt{n\alpha_n}\right) \geq 1 - n^{-1}/2\,.$$

Proof. Observe that $\|A^{(1)} - P^{(1)}\| \leq \|A - P\|$. The claimed result follows easily from Theorem 5.2 of Lei & Rinaldo (2013), where it has been shown that with high probability $\|A - P\| \leq C\sqrt{n\alpha_n}$. ∎

Lemma 5 (Singular subspace error bound)**.** Let $\hat{M}$, M be two matrices of same dimension, and $\hat{U}$ and U be $n \times K$ orthonormal matrices corresponding to the top K right singular vectors of $\hat{M}$ and M, respectively. Then there exists a $K \times K$ orthogonal matrix Q such that

$$\|\hat{U} - UQ\| \leq \frac{2\sqrt{2}\|\hat{M} - M\|}{\sigma_K(M)}\,.$$

Proof. If $\|\hat{M} - M\| \leq \sigma_K(M)/2$, then using Wedin $\sin\Theta$ theorem (Wedin, 1972) and Weyl's inequality there exists an orthogonal Q such that $\|\hat{U} - UQ\| \leq \|\hat{M} - M\|/(\sigma_K(M) - \|\hat{M} - M\|) \leq 2\|\hat{M} - M\|/\sigma_K(M)$. If $\|\hat{M} - M\| \geq \sigma_K(M)/2$, then $\|\hat{U} - UQ\| \leq 1 \leq 2\|\hat{M} - M\|/\sigma_K(M)$. ∎

Remark. The orthogonal matrix Q will have no particular impact on the argument below. For presentation simplicity, we assume, without loss of generality, that $Q = I$ in the rest of the proof.

Since our community recovery method is spectral clustering, the proof of Theorem 1 relies on the following lemma that guarantees the accuracy of k-means algorithm under a Frobenius norm condition.

Lemma 6 (Lemma 5.3 of Lei & Rinaldo (2013))**.** Let $\hat{U}$ and U be two $n \times K$ matrices such that U contains K distinct rows. Let δ be the minimum distance between two distinct rows of U, and g be the membership vector given by clustering the rows of U satisfying Assumption A2. Let $\hat{g}$ be the output of a k-means clustering algorithm on $\hat{U}$, with objective value no larger than a constant factor of the global optimum. Assume that $\|\hat{U} - U\|_F^2 \leq cn\delta^2$ for some small enough constant c. Then $\hat{g}$ agrees with g on all but $c^{-1}\|\hat{U} - U\|_F^2\delta^{-2}$ nodes after an appropriate label permutation.

Proof of Theorem 1. Let G be the $n \times K$ matrix with $G_{ij} = 1$ if $j = g_i$, and $G_{ij} = 0$ otherwise. Let $G^{(1)}$ be the submatrix of G containing rows in $\mathcal{N}_1$. Then $P^{(1)} = G^{(1)}BG^T = G\tilde{B}\tilde{G}^T$, where $\tilde{G}$ is an $n \times K$ matrix obtained by normalizing the columns of G. and $\tilde{B}$ is

a $K \times K$ matrix after corresponding column scaling of B. It is easy to check that $\tilde{G}$ has orthonormal columns and hence the top K-dimensional right singular subspace of $P^{(1)}$ is spanned by $U = \tilde{G}Q$, for any $K \times K$ orthogonal matrix Q. Thus U contains K distinct rows and the distance between two distinct rows is of order at least $1/\sqrt{n}$ under Assumption A2.

We focus on the event that $\min_k n^*_{1,k} \geq c_0\pi_0 n/2$ and $\|A^{(1)} - P^{(1)}\| \leq C\sqrt{n\alpha_n}$, which has probability at least $1 - n^{-1}$ according to Lemmas 3 and 4. Then it can be directly verified that $\sigma_K(P^{(1)}) \geq Cn\alpha_n$ because the Kth largest singular values of both $G^{(1)}$ and G are of order at least $\sqrt{n}$. Then Lemma 5 implies that $\hat{U}$ and U, the matrices consisting of $n \times K$ top singular vectors of $A^{(1)}$ and $P^{(1)}$, satisfy, with appropriate choice of Q,

$$\|\hat{U} - U\|_F^2 \leq C\frac{1}{n\alpha_n}.$$

Then applying Lemma 6, we know that the k-means clustering algorithm misclusters no more than C/α_n nodes. This completes the proof of part (a).

To prove part (b), for fixed $1 \leq k < k' \leq K$, consider the oracle estimator

$$\hat{B}^*_{k,k'} = \frac{\sum_{i\in\mathcal{N}^*_{1,k}, j\in\mathcal{N}^*_{1,k}\cup\mathcal{N}^*_{2,k'}} A_{ij}}{n^*_{1,k}(n^*_{1,k'} + n^*_{2,k'})}. \tag{5}$$

It is obvious that $\hat{B}^*_{k,k'} = B_{k,k'}(1 + o_P(1))$ because $n^*_{j,k} \to \infty$ at order n for all j, k.

Now using part (a), the numerators of (3) and (5) differ by at most $o(n^{3/2})$. But the denominators are asymptotically equivalent with order n^2 and their ratio tends to 1. Thus $|\hat{B}_{k,k'} - \hat{B}'_{k,k'}| = o_P(n^{-1/2}) = o_P(\hat{B}_{k,k'})$. The proof for $k = k'$ is almost identical. ∎

For the proof of Theorem 2, we need the analogous version of Lemma 6 for the k-median algorithm, which is a simple adaptation of Lemma 6 and has been proved in Theorem 4.2 of Lei & Rinaldo (2013). For a matrix M, $\|M\|_{2,1}$ denotes the sum of ℓ_2 norms of the rows in M.

Lemma 7. Let $\hat{U}$ and U be two $n \times K$ matrices such that U contains K distinct rows. Let δ be the minimum distance between two distinct rows of U, and g be the membership vector given by clustering the rows of U. Let $\hat{g}$ be the output of a k-median clustering algorithm on $\hat{U}$, with objective value no larger than a constant factor of the global optimum. Assume that $\|\hat{U} - U\|_{2,1} \leq cn\delta$ for some small enough constant c and that g satisfies Assumption A2. Then $\hat{g}$ agrees with g on all but $c^{-1}\|\hat{U} - U\|_{2,1}\delta^{-1}$ nodes after an appropriate label permutation.

Proof of Theorem 2. Let Ψ be an $n \times K$ matrix such that $\Psi_{ij} = \psi'_i$ if $j = g_i$ and $\Psi_{ij} = 0$ otherwise. Let $\Psi^{(1)}$ be the corresponding submatrix of Ψ with rows in $\mathcal{N}_1$. Then $P^{(1)} =$

$\Psi^{(1)}\tilde{B}\Psi$, where $\tilde{B}$ is a $K \times K$ matrix obtained after corresponding row/column scaling of B. It is easy to check that Ψ is orthonormal so that the top K-dimensional right singular subspace of $P^{(1)}$ is spanned by $U = \Psi Q$ for any $K \times K$ orthogonal Q. It follows that the norm of ith row of U is ψ'_i, and that any two rows of U in distinct communities are orthogonal. Let $\tilde{U}$ and $\tilde{U}^*$ be the row-normalized versions of $\hat{U}$ and U, respectively. Then $\tilde{U}^*$ contains K distinct rows and the distance between any two distinct rows of $\tilde{U}^*$ is $\sqrt{2}$.

Similarly we will focus on the event that $\min_k n^*_{1,k} \geq c_0\pi_0 n/2$ and $\|A^{(1)} - P^{(1)}\| \leq C\sqrt{n\alpha_n}$, which has probability at least $1 - n^{-1}$ according to Lemmas 3 and 4. Using the same reasoning as in the proof of Theorem 1 we know that, for appropriate choice of Q,

$$\|\hat{U} - U\|_F^2 \leq C/(n\alpha_n)\,.$$

Because the minimum row norm of U is at least $\psi_0/\sqrt{n}$, the number of zero rows in $\hat{U}$ is at most $\|\hat{U} - U\|_F^2/(\psi_0/\sqrt{n})^2 = O(\alpha_n^{-1}) = o(\sqrt{n/\alpha_n})$. In the rest of the proof we can safely assume that $\hat{U}$ has no zero rows.

Now let u_i, $\hat{u}_i$ be the ith row of U, $\hat{U}$, respectively. We have, using the fact that $\|(u/\|u\| - v/\|v\|)\| \leq 2\|u-v\|/\|v\|$ for all vectors u, v of same dimension, Cauchy-Schwartz, and Assumption A4,

$$\begin{aligned}\|\tilde{U} - \tilde{U}^*\|_{2,1} \leq & 2\sum_{i=1}^n \frac{\|\hat{u}_i - u_i\|}{\|u_i\|} = 2\sum_{i=1}^n \frac{\|\hat{u}_i - u_i\|}{\psi'_i} \\ \leq & 2\|\hat{U} - U\|_F \left(\sum_{i=1}^n (\psi'_i)^{-2}\right)^{1/2} \leq 2\psi_0^{-1} n\|\hat{U} - U\|_F \leq C\sqrt{n/\alpha_n}\,.\end{aligned}$$

Then part (a) follows by applying Lemma 7 to $\tilde{U}$ and $\tilde{U}^*$.

For part (b), recall that $\|u_i\| = \psi'_i$ for all i. Then Cauchy-Schwartz implies that

$$\|\hat{\psi}' - \psi'\|_1 \leq \sum_{i=1}^n \|\hat{u}_i - u_i\| \leq \sqrt{n}\|\hat{U} - U\|_F \leq C\alpha_n^{-1/2}\,. \tag{6}$$

By Assumptions (A2) and (A4) we have $\inf_i \psi'_i \geq Cn^{-1/2}$ for some constant C.

Let $S_n = \{i : |\hat{\psi}'_i - \psi'_i| \leq n^{-1/2}(n^{1/3}\alpha_n)^{-1/2}\}$. Then for all $i \in S_n$, we have $\hat{\psi}'_i = \psi'_i(1 + o(1))$ and

$$|S_n^c| \leq \frac{\|\hat{\psi}' - \psi'\|_1}{n^{-2/3}\alpha_n^{-1/2}} \leq Cn^{2/3}\,. \tag{7}$$

For $1 \leq k < k' \leq K$, consider the oracle estimator

$$\hat{B}^*_{k,k'} = \frac{\sum_{i\in\mathcal{N}^*_{1,k}, j\in\mathcal{N}^*_{1,k'}\cup\mathcal{N}^*_{2,k'}} A_{ij}}{\sum_{i\in\mathcal{N}^*_{1,k}, j\in\mathcal{N}^*_{1,k'}\cup\mathcal{N}^*_{2,k'}} \psi'_i\psi'_j}\,.$$

It is obvious that $\hat{P}^*_{ij} = \psi'_i\psi'_j\hat{B}'^*_{g_ig_j} = (1+o_P(1))\psi_i\psi_j B_{g_ig_j} = (1+o_P(1))P_{ij}$. As a result, the claim in part (b) of the theorem follows if we can show that $\hat{B}'_{k,k'} = (1+o_P(1))\hat{B}'^*_{k,k'}$ because for all but a vanishing proportion of pairs (i,j) we have $(\hat{g}_i,\hat{g}_j) = (g_i,g_j)$ and $\hat{\psi}'_i\hat{\psi}'_j = \psi'_i\psi'_j(1+o(1))$ in view of (7). To this end, we compare

$$n^{-1}\hat{B}'_{k,k'} = \frac{\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}^*_{2,k'}} A_{ij}}{\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} (\sqrt{n}\hat{\psi}'_i)(\sqrt{n}\hat{\psi}'_j)}$$

with

$$n^{-1}\hat{B}'^*_{k,k'} = \frac{\sum_{i\in\mathcal{N}^*_{1,k}, j\in\mathcal{N}^*_{1,k'}\cup\mathcal{N}^*_{2,k'}} A_{ij}}{\sum_{i\in\mathcal{N}^*_{1,k}, j\in\mathcal{N}^*_{1,k'}\cup\mathcal{N}^*_{2,k'}} (\sqrt{n}\psi'_i)(\sqrt{n}\psi'_j)}\,.$$

Note that $\sqrt{n}\psi'_i \asymp 1$ for all i. It is easy to check that the numerators differ by at most $o(n^{5/3})$. For the denominators, first we compare the denominator of $n^{-1}\hat{B}'^*_{k,k'}$ with

$$\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} (\sqrt{n}\psi'_i)(\sqrt{n}\psi'_j)\,. \tag{8}$$

It straightforward to check that their ratio tends to 1, because the index sets of these two summations differ by a vanishing proportion. Now compare (8) with the denominator of $n^{-1}\hat{B}'_{k,k'}$. We have

$$\begin{aligned}
\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} (\sqrt{n}\hat{\psi}'_i)(\sqrt{n}\hat{\psi}'_j) &= \left(\sum_{i\in\mathcal{N}_{1,k}} \sqrt{n}\hat{\psi}'_i\right)\left(\sum_{j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} \sqrt{n}\hat{\psi}'_j\right)\\
&= (1+o(1))\left(\sum_{i\in\mathcal{N}_{1,k}} \sqrt{n}\psi'_i\right)\left(\sum_{j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} \sqrt{n}\psi'_j\right)\\
&= (1+o(1)) \sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} (\sqrt{n}\psi'_i)(\sqrt{n}\psi'_j)\,,
\end{aligned}$$

where the second line follows from the fact $\sum_{i\in\mathcal{N}_{j,k}} \hat{\psi}'_i = (1+o(1))\sum_{i\in\mathcal{N}_{j,k}} \psi'_i$ for all $j\in\{1,2\}$, $1\le k\le K$, which is a consequence of (6). Now we conclude that the denominators of $n^{-1}\hat{B}'_{k,k'}$ and $n^{-1}\hat{B}'^*_{k,k'}$ are both of order at least n^2 with ratio tending to one. Therefore, the absolute difference between $n^{-1}\hat{B}'_{k,k'}$ and $n^{-1}\hat{B}'^*_{k,k'}$ is $o(n^{-1/3})$ which is $o(n^{-1}\hat{B}'^*_{k,k'})$. The same argument can be used for the case $k = k'$. ∎

References

Abbe, E., Bandeira, A. S., & Hall, G. (2014). Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*.

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, (pp. 36–43). ACM.

Anandkumar, A., Ge, R., Hsu, D., & Kakade, S. M. (2014). A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, *15*, 2239–2312.

Bickel, P. J., & Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068–21073.

Bickel, P. J., & Sarkar, P. (2013). Hypothesis testing for automated community detection in networks. *arXiv preprint arXiv:1311.2694*.

Charikar, M., Guha, S., Tardos, É., & Shmoys, D. B. (1999). A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, (pp. 1–10). ACM.

Chaudhuri, K., Chung, F., & Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *JMLR: Workshop and Conference Proceedings*, *2012*, 35.1–35.23.

Chen, Y., Sanghavi, S., & Xu, H. (2012). Clustering sparse graphs. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.) *Advances in Neural Information Processing Systems 25*, (pp. 2213–2221).

Decelle, A., Krzakala, F., Moore, C., & Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, *84*(6), 066106.

Faust, K., & Wasserman, S. (1992). Blockmodels: Interpretation and evaluation. *Social networks*, *14*(1), 5–61.

Fishkind, D. E., Sussman, D. L., Tang, M., Vogelstein, J. T., & Priebe, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, *34*(1), 23–39.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, *97*(460), 1090–1098.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, *5*(2), 109–137.

Jin, J. (2012). Fast community detection by SCORE. arXiv:1211.5803.

Karrer, B., & Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, *83*(1), 016107.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI*, vol. 3, (p. 5).

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., & Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, *110*(52), 20935–20940.

Lei, J., & Rinaldo, A. (2013). Consistency of spectral clustering in sparse stochastic block models. *arXiv preprint arXiv:1312.2050*.

Li, S., & Svensson, O. (2013). Approximating k-median via pseudo-approximation. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, (pp. 901–910). ACM.

Massoulie, L. (2013). Community detection thresholds and the weak ramanujan property. *arXiv preprint arXiv:1311.3085*.

McSherry, F. (2001). Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, (pp. 529–537). IEEE.

Mossel, E., Neeman, J., & Sly, A. (2013). A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Skala, M. (2013). Hypergeometric tail inequalities: ending the insanity. *arXiv preprint arXiv:1311.5939*.

Sussman, D. L., Tang, M., & Priebe, C. E. (2013). Universally consistent latent position estimation and vertex classification for random dot product graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *36*(1), 48–57.

Vu, V. (2014). A simple svd algorithm for finding hidden partitions. *arXiv preprint arXiv:1404.3918*.

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, *12*(1), 99–111.

Yan, X., Shalizi, C., Jensen, J. E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P., & Zhu, Y. (2014). Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, *2014*(5), P05007.

Young, S. J., & Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *Algorithms and models for the web-graph*, (pp. 138–149). Springer.

Zhao, Y., Levina, E., & Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences*, *108*(18), 7321–7326.

Zhao, Y., Levina, E., & Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, *40*(4), 2266–2292.