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

Kehui Chen and Jing Lei
University of Pittsburgh and Carnegie Mellon University

August 1, 2016

Abstract

The stochastic block model and its variants have been a popular tool for analyzing large network data with community structures. In this paper we develop an efficient *network cross-validation* (NCV) approach to determine the number of communities, as well as to choose between the regular stochastic block model and the degree corrected block model. The proposed NCV method is based on a *block-wise node-pair splitting* technique, combined with an integrated step of community recovery using sub-blocks of the adjacency matrix. We prove that the probability of under-selection vanishes as the number of node increases, under mild conditions satisfied by a wide range of popular community recovery algorithms. The solid performance of our method is also demonstrated in extensive simulations and two data examples.

KEYWORDS: stochastic block models, community recovery, model selection, cross-validation, block-wise node-pair splitting

---

Kehui Chen is Assistant Professor, Department of Statistics and Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15260 (E-mail: khchen@pitt.edu). Jing Lei is Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: jinglei@andrew.cmu.edu). Jing Lei's research is partially supported by NSF Grant DMS-1407771.

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. 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 and Girvan, 2004; Newman, 2006). The stochastic block model (Holland et al., 1983) and its variants, such as the degree corrected block model (Karrer and Newman, 2011), are powerful and mathematically elegant tools for modeling large networks with community structures, and have been proved useful in many scientific areas such as social science, biology, and information science (Faust and Wasserman, 1992; Kemp et al., 2006; Bickel and Chen, 2009).

The community recovery problem for stochastic block models has been the focus of much research effort in several areas including statistics (Daudin et al., 2008; Bickel and Chen, 2009; Zhao et al., 2012; Jin, 2015; Latouche et al., 2012; Fishkind et al., 2013; Lei and Rinaldo, 2015), 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 (Massoulié, 2014; Mossel et al., 2013; Abbe et al., 2016). 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 community recovery, determining the number of communities has remained a challenging problem and has gained much interest recently. In principle, one can transform the problem to selecting the number of clusters in a regular multivariate data using techniques such as spectral embedding (Sussman et al., 2012). But this would involve choosing a good embedding space with appropriate dimensionality, which itself is a challenging task. 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 and Sarkar (2016) 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 a

suitably scaled and centered adjacency matrix. But the convergence rate is slow and a bootstrap correction is needed in practice. Moreover, these sequential or recursive testing procedures only work for certain types of community structures. After the first draft of this work, there have been some new developments along the line of testing $K = \widetilde{K}$ for a given candidate value $\widetilde{K}$ (Lei, 2016). Some other model selection criteria have also been proposed, using various techniques including likelihood-based methods (Handcock et al., 2007; Daudin et al., 2008; Airoldi et al., 2008; Saldana et al., 2014; Wang and Bickel, 2015), Bayesian inference (Latouche et al., 2012; McDaid et al., 2013), and information theory (Rosvall and Bergstrom, 2007; Peixoto, 2013).

In this paper, we focus on a generic idea of network cross-validation. Cross-validation is a very popular and appealing method in many model selection problems. The adaptation to network data is usually through a node splitting procedure and has been considered by Airoldi et al. (2008), among others. A random node-pair splitting method has been used in Hoff (2008) for model selection under a Bayesian framework. These methods, even though applicable to the community recovery problem, are usually computationally intensive. Moreover, the theoretical investigation for cross-validation methods in network model selection remains open.

The network cross-validation (NCV) method developed in this paper is based on a *block-wise node-pair splitting* technique. Simply speaking, the splitting step divides the nodes randomly into two groups $\mathcal{N}_1$ and $\mathcal{N}_2$. The observed edge formation between node pairs $\{(i,j) : i \in \mathcal{N}_1, j \in \mathcal{N}_1 \cup \mathcal{N}_2\}$ are used as the fitting set, and the node pairs $\{(i,j) : i,j \in \mathcal{N}_2\}$ are used as the testing set. Such a node-pair splitting is superior to a simple node splitting. It originates from two key observations. First, the fitting set carries full information about the network model parameters. That is, we can consistently estimate, using only data in the fitting set, the membership of all the nodes as well as the community-wise edge probability matrix. Second, given the community membership, the data in the fitting set and in the testing set are independent. The second observation reflects a significant difference between the traditional parameterization of the stochastic block model that treats the node memberships as missing 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 and Chen, 2009). However, for community recovery based on a single observed realization of the stochastic block model, the useful

information for statistical inference is largely contained in the randomness of edge formation. In contrast, by treating the membership as parameters, the conditional parameterization substantially simplifies the relationship between the distributions of edge formation in the fitting set and in the testing set.

We describe the algorithm in detail in Section 2. The proposed V-fold NCV method is novel and is of substantial practical interest for several reasons. First, it is computationally efficient, requiring only one model fitting for each fold. Second, it is tuning free except the number of folds. Moreover, it is general enough to be combined with different community recovery techniques. In Section 3, we characterize the theoretical properties of the proposed NCV method. We show that under appropriate conditions, when combined with popular community recovery techniques, such as modularity based optimization and spectral clustering, the proposed NCV does not underestimate the number of communities with probability tending to one. Protection against overestimation is also discussed. In Section 5, we demonstrate the effectiveness of our method via extensive numerical experiments, where different types of network community structures are investigated.

The NCV method can be applied to select the best model from a general collection of candidate models, which does 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 node-pair splitting idea behind NCV can be further extended to other network models with conditional edge independence. These extensions are described in Section 4 and Section 6, and are illustrated in an application to a political blog data in Section 5, where the NCV method chooses the degree corrected block model with two communities, matching pervious findings in the literature.

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 an $n \times n$ symmetric binary adjacency matrix A. The community structure is represented by a vector $g = (g_1, \ldots, g_n)$ with $g_i \in \{1, ..., K\}$ 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

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

where B is a $K \times K$ symmetric matrix representing the community-wise edge probabilities. 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 will be discussed in later sections.

We first describe the overall algorithm of the V-fold network cross-validation procedure. Each step will be explained in details for a single fold in subsequent subsections.

Algorithm 1: 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. *Block-wise node-pair splitting:*

   Randomly split the nodes into V equal-sized subsets $\{\widetilde{\mathcal{N}}_v : 1 \leq v \leq V\}$, and split the adjacency matrix correspondingly into $V \times V$ equal sized blocks

   $$A = (\widetilde{A}^{(uv)} : 1 \leq u, v \leq V), \tag{2}$$

   where $\widetilde{A}^{(uv)}$ is the submatrix of A with rows in $\widetilde{\mathcal{N}}_u$ and columns in $\widetilde{\mathcal{N}}_v$.

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

   (a) *Estimation from the rectangular matrix:*

   Estimate model parameters $(\widehat{g}^{(v)}, \widehat{B}^{(v)})$ using the rectangular submatrix obtained by removing the rows of A in subset $\widetilde{\mathcal{N}}_v$

   $$\widetilde{A}^{(-v)} = (\widetilde{A}^{(rs)} : r \neq v,\ 1 \leq r, s \leq V)\,.$$

   (b) *Validation:*

   Calculate the predictive loss evaluated on $\widetilde{A}^{(vv)}$:

   $$\widehat{L}^{(v)}(A, \widetilde{K}) = \sum_{i,j \in \widetilde{\mathcal{N}}_v, i \neq j} \ell(A_{ij}, \widehat{P}^{(v)}_{ij})\,, \tag{3}$$

   where $\widehat{P}^{(v)}_{ij} = \widehat{B}^{(v)}_{\widehat{g}^{(v)}_i, \widehat{g}^{(v)}_j}$.

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

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

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 6. There is some randomness in $\widehat{K}$ due to the randomness in V-fold splitting. We recommend to perform NCV multiple times with independent V-fold splits and output the most frequent value of $\widehat{K}$. In our simulation, twenty repetitions was sufficient to stabilize the output.

2.1 Step 1: block-wise node-pair splitting

Now we further discuss the block-wise node-pair splitting for a single fold v. Recall that in the V-fold split of adjacency matrix A, we denote $\widetilde{A}^{(rs)}$ the submatrix of A that corresponds to the rows in the rth fold and columns in the sth fold. When fold v is used for validation, we only need to consider nodes inside and outside of the vth fold. To emphasize this point and facilitate discussion, let $A^{(11)} = (\widetilde{A}^{(rs)} : r \neq v, s \neq v)$, $A^{(22)} = \widetilde{A}^{(vv)}$, $A^{(12)} = (\widetilde{A}^{(rv)} : r \neq v)$. We can re-arrange the $V \times V$ block adjacency matrix in a collapsed 2×2 block form.

$$A = \begin{pmatrix} A^{(11)} & A^{(12)} \\ A^{(21)} & A^{(22)} \end{pmatrix}, \tag{4}$$

Such a splitting puts node pairs in $A^{(11)}$ and $A^{(12)}$ as the fitting sample and those in $A^{(22)}$ as the validating sample. Finally for ease of presentation we use $\mathcal{N}_1$ to denote $\widetilde{\mathcal{N}}_v^c$ and $\mathcal{N}_2$ for $\widetilde{\mathcal{N}}_v$.

The block-wise node-pair splitting arranges the fitting sample and validating sample in a block matrix form. An important consequence is that one can accurately estimate all model parameters, including the community membership of all nodes, from the fitting sample. Such a splitting method has several appealing features when compared with existing cross-validation methods for network data that are mostly based on a node splitting technique. In the node splitting method, where the nodes, instead of the node pairs, are split into a fitting set and a testing set, one typically assumes that the node memberships are generated independently with probability $P(g_i = k) = \pi_k$ for some $\pi = (\pi_1, ..., \pi_K)$ such that $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$. After the node splitting, the model parameter (π, B) is estimated on the subnetwork confined on the fitting set of nodes, and evaluated on the subnetwork confined on the testing subset of nodes. This approach has some drawbacks. First, calculating the full likelihood in terms of π and B 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, which often leads to inaccurate model selection. Third, it does not use the observed edge formation between the fitting and testing nodes. The first two drawbacks are also present in the regular BIC approach using full likelihood, as pointed out by Handcock et al. (2007); Airoldi et al. (2008). They proposed a hybrid criterion, which combines a conditional log-likelihood given an estimated membership, with a BIC-type penalty term based on the unconditional parameterization. In the current paper, the NCV framework fully takes a conditional perspective, where the node memberships are treated as fixed and estimated from the fitting submatrix, which makes NCV fundamentally different from the regular cross-validation based on node splitting. In addition, NCV fully exploits the information carried in node pairs between $\widetilde{\mathcal{N}}_v$ and $\widetilde{\mathcal{N}}_v^c$, making the model fitting and validation statistically and computationally more efficient.

2.2 Step 2: estimating model parameters from the rectangular matrix

After splitting, we estimate model parameters (g, B) from the $n_1 \times n$ rectangular matrix $A^{(1)} = (A^{(11)}, A^{(12)})$. Many standard community recovery procedures designed for the full adjacency matrix can be extended to this case. In our numerical examples we have implemented spectral clustering and a least squares estimator. We first illustrate the implementation of spectral clustering.

For a given candidate value $\widetilde{K}$ of K, 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 $n \times \widetilde{K}$ matrix consisting of the leading d right singular vectors.

Rectangular spectral clustering:

Input: Rectangular $n_1 \times n$ matrix $A^{(1)}$, a candidate number of communities $\widetilde{K}$, number of spectral components d (default choice of d is $\widetilde{K}$)

1. Let $\widehat{U}$ be the $n \times d$ matrix consisting of the top d right singular vectors of $A^{(1)}$.
2. Output $\widehat{g}$ by applying the k-means clustering algorithm with $\widetilde{K}$ clusters to the rows of $\widehat{U}$.

The selection of d in spectral clustering concerns the problem of choosing the number of significant singular values of the adjacency matrix (Owen and Perry, 2009; Josse and Husson, 2012; Chatterjee, 2014). In a stochastic block model, d also corresponds to the rank of the community-wise edge probability matrix B. Choosing d is not necessarily the same as choosing the number of communities (Fishkind et al., 2013; Lei, 2016). In practice, one may choose d by looking at the

scree plot of singular values of A. In our simulation studies, we use $d = \widetilde{K}$ according to the fact that in a stochastic block model with $\widetilde{K}$ communities the rank of B does not exceed $\widetilde{K}$. Such a choice may include more singular vectors than necessary, but the number of redundant singular vectors will not be large for moderate values of $\widetilde{K}$.

Next we describe the least squares estimator, which can be written as

$$\widehat{g}_{\text{ls}} = \arg \max_{g \in \{1,\ldots,\widetilde{K}\}^n} \max_{B \in [0,1]^{\widetilde{K} \times \widetilde{K}}} \sum_{1 \le i < j \le n} (A_{ij} - B_{g_i g_j})^2 .$$

The least squares estimator has been studied in the network data literature (Gao et al., 2015; Borgs et al., 2015) and can be viewed as a variant of the likelihood estimator (Bickel and Chen, 2009)

$$\widehat{g}_{\text{mle}} = \arg \max_{g \in \{1,\ldots,\widetilde{K}\}^n} \max_{B \in [0,1]^{\widetilde{K} \times \widetilde{K}}} \left(\prod_{i,j \in \mathcal{N}_1, i<j} B_{g_i g_j}^{A_{ij}} (1 - B_{g_i g_j})^{1 - A_{ij}} \right) \times \left(\prod_{i \in \mathcal{N}_1, j \in \mathcal{N}_2} B_{g_i g_j}^{A_{ij}} (1 - B_{g_i g_j})^{1 - A_{ij}} \right) .$$

In practice, both maximization problems can be solved using heuristic search methods such as the label switching algorithm (Bickel and Chen, 2009) and the tabu search (Zhao et al., 2012). Empirically these two estimators have similar performance. In this paper we focus on the least squares method which is slightly more stable when the network is sparse.

Once $\widehat{g}$ is obtained, let $\mathcal{N}_{j,k}$ be the nodes in $\mathcal{N}_j$ with estimated membership k, and $n_{j,k} = |\mathcal{N}_{j,k}|$ ($j = 1, 2$, $1 \le k \le \widetilde{K}$). We can estimate B using a simple plug-in estimator:

$$\widehat{B}_{k,k'} = \begin{cases} \dfrac{\sum_{i \in \mathcal{N}_{1,k}, j \in \mathcal{N}_{1,k'} \cup \mathcal{N}_{2,k'}} A_{ij}}{n_{1,k}(n_{1,k'} + n_{2,k'})} , & k \neq k' , \\ \dfrac{\sum_{i,j \in \mathcal{N}_{1,k}, i<j} A_{ij} + \sum_{i \in \mathcal{N}_{1,k}, j \in \mathcal{N}_{2,k}} A_{ij}}{(n_{1,k} - 1) n_{1,k}/2 + n_{1,k} n_{2,k}} , & k = k' . \end{cases} \tag{5}$$

We provide theoretical analysis of these estimators in Section 3. The success of these estimators requires adequate presence of all communities in the training subsample. Thus the NCV method typically works well when the communities are balanced, a condition typically holds when the number of communities is not too large. This is reflected in our simulation studies in Section 5.

2.3 Step 3: validation using the testing set

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

For each observation 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 estimated by $\widehat{P}_{ij} = \widehat{B}_{\widehat{g}_i \widehat{g}_j}$. Some natural choices of the loss function ℓ in (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.

In the validation step, if the candidate value $\widetilde{K}$ 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 $\widetilde{K}$ is too large, then the model overfits the data, with noisy prediction on the testing data. Therefore, it is natural to expect the validated predictive loss $\widehat{L}(A, \widetilde{K}) = \sum_{i,j \in \mathcal{N}_2, i \neq j} \ell(A_{ij}, \widehat{P}_{ij})$, to be minimized when $\widetilde{K} = K$, the true number of communities. Partial theoretical supports and further heuristic arguments are given in the following section.

3. THEORETICAL PROPERTIES FOR NCV IN STOCHASTIC BLOCK MODELS

For two sequences of positive numbers a_n and b_n, we denote $a_n = \Omega(b_n)$ if $\liminf_{n \to \infty} a_n / b_n > 0$, and $a_n = \omega(b_n)$ if $a_n / b_n \to \infty$. We also adopt the standard notation where $a_n = O(b_n)$ means that $\limsup_{n \to \infty} a_n / b_n < \infty$ and $a_n = o(b_n)$ means that $a_n / b_n \to 0$.

To study the asymptotic behavior of the estimator, we consider a sequence of SBM's parameterized by $(g^{(n)}, B^{(n)} : n \geq 1)$ such that

(A1) $g^{(n)}$ is a membership vector of length n with K distinct communities, and the minimum community size is at least $\pi_0 n$ for some constant π_0.

(A2) $B^{(n)} = \rho_n B_0$ where B_0 is a $K \times K$ symmetric matrix with entries in $(0, 1]$, and the rows of B_0 are all distinct. The rate ρ_n, which controls the network sparsity, satisfies $\rho_n = \Omega(\log n / n)$.

The first condition requires the block sizes to be relatively balanced. It is satisfied with high probability if the memberships are independently generated from a multinomial distribution. The second condition allows the edge probability to decrease at a rate ρ_n as n increases. The lower bound on the sparsity ρ_n makes it possible to obtain accurate community recovery.

Community recovery is an integrated part in the proposed NCV method and its accuracy plays an important role in the performance of model selection. We introduce two notions of community recovery consistency.

Definition 1 (Exactly consistent recovery). *Given a sequence of SBM's with K blocks parameterized by $(g^{(n)}, B^{(n)})$, we call a community recovery method $\widehat{g}$* exactly consistent *if $P(\widehat{g}(A, K) = g^{(n)}) \to 1$, where A is a realization of SBM $(g^{(n)}, B^{(n)})$ and the equality is up to a possible label permutation.*

Definition 2 (Approximately consistent recovery). *For a sequence of SBM's with K blocks parameterized by $(g^{(n)}, B^{(n)})$ and a sequence $\eta_n = o(1)$, we say $\widehat{g}$ is approximately consistent with rate η_n if,*

$$\lim_{n\to\infty} P[\mathrm{Ham}(\widehat{g}(A, K), g) \geq \eta_n n] = 0,$$

where $\mathrm{Ham}(\widehat{g}, g)$ is the smallest Hamming distance between $\widehat{g}$ and g among all possible label permutations.

Exactly consistent community recovery can be achieved under mild assumptions on $g^{(n)}$ and $B^{(n)}$. It is known that likelihood methods (Bickel and Chen, 2009) are exactly consistent when $\rho_n n/\log n \to \infty$, and variants of spectral methods (McSherry, 2001; Vu, 2014; Lei and Zhu, 2014) are exactly consistent when $\rho_n n/\log n > C$ for some constant C depending on π_0, B_0 only. Consistency of the least squares estimator are provided in Gao et al. (2015); Borgs et al. (2015). The proposed NCV method can be combined with any of these community recovery methods. For the simple spectral clustering algorithm, the exact consistency result has not been established. But its highly competitive empirical performance has been observed in extensive numerical studies and it is popular due to its simple implementation. In the following, we first establish the approximately consistent recovery result for the simple spectral clustering algorithm described in 2.2.

Theorem 1 (Consistency of spectral clustering). *Assuming that B_0 is non-singular and A1-A2 hold, for each fold split in NCV, the $\widehat{g}$ estimated using spectral clustering as described in Section 2.2 is approximately consistent with rate $(n\rho_n)^{-1}$.*

Now we state the main theorem. For a given V-fold block-wise partition of A, recall that $\widetilde{\mathcal{N}}_v$ is the set of nodes in the vth fold. Let

$$L(A) = \sum_{v=1}^{V} \sum_{i,j\in\widetilde{\mathcal{N}}_v, i\neq j} \ell(A_{ij}, B_{g_i g_j}) .$$

Theorem 2. *Under conditions A1–A2, with the loss function* $\ell(a,p) = (a-p)^2$*, for a candidate* $\widetilde{K}$*, we have*

(a) When $\widetilde{K} < K$*, for* any *estimator* $(\widehat{g}, \widehat{B})$*, we have*

$$\widehat{L}(A, \widetilde{K}) - L(A) \geq c(n\rho_n)^2 + O_P(1)\,.$$

(b) When $\widetilde{K} = K$*, if* $\widehat{g}$ *is exactly consistent and* $\widehat{B}$ *is estimated as in (5), then*

$$\left|\widehat{L}(A, \widetilde{K}) - L(A)\right| = O_P(n\rho_n^{3/2}).$$

(c) When $\widetilde{K} = K$*, if* $\widehat{g}$ *is approximately consistent with rate* η_n *and* $\widehat{B}$ *is estimated as in (5), then*

$$\left|\widehat{L}(A, \widetilde{K}) - L(A)\right| = O_P(n\rho_n^{3/2} + n^2\rho_n\eta_n)\,.$$

As a consequence, when the candidate set $\mathcal{K}$ is fixed and contains the truth, we have the following guarantee against under selection.

Corollary 3. *Assume that the candidate set* $\mathcal{K}$ *is fixed, independent of* n*, and contains* K*. Assume also that conditions A1–A2 hold. Consider loss function* $\ell(a,p) = (a-p)^2$*.*

(a) When $\widehat{g}$ *is exactly consistent and* $\widehat{B}$ *is estimated as in (5), we have*

$$\lim_{n\to\infty} P(\widehat{K} < K) = 0$$

(b) When $\widehat{g}$ *is obtained by the simple spectral clustering method described in Section 2.2, and* $\widehat{B}$ *is estimated as in (5), we have*

$$\lim_{n\to\infty} P(\widehat{K} < K) = 0\,,$$

provided that $\rho_n = \omega(n^{-1/2})$ *and* B_0 *is non-singular.*

The proofs of Theorem 1 and Theorem 2 are given in Appendix A. Part (a) of Corollary 3 is a direct consequence of part (a) and (b) of Theorem 2. Part (b) of Corollary 3 can be easily derived by combining Theorem 1 and part (a), (c) of Theorem 2.

As commonly known for cross-validation methods, there is no corresponding theoretical guarantee against overestimation. Here we provide some heuristic understanding why the NCV method

will tend not to give $\widehat{K} > K$ in stochastic block models. The estimated communities must belong to one of two scenarios. In the first scenario, each of the estimated community contains mostly nodes from one true community. So if $\widetilde{K} > K$, at least one true community is artificially split, and the corresponding $B_{k,k'}$ is estimated separately in these artificially split blocks. The difference between these separate estimates mainly reflects spurious random fluctuations, which will only lead to a larger predictive loss on an independent testing data. In the second scenario, at least one estimated community contains a substantial proportion of nodes from two true communities. In this case, the cross-validated predictive loss can be shown to be large regardless of $\widetilde{K} > K$ or $\widetilde{K} < K$, using the same argument as in part (a) of Theorem 2.

4. DEGREE CORRECTED BLOCK MODELS AND FURTHER EXTENSIONS

4.1 Choosing K for degree corrected block models

The degree corrected block model (Karrer and 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

$$P(A_{ij} = 1) = 1 - P(A_{ij} = 0) = \psi_i \psi_j B_{g_i g_j}, \tag{6}$$

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. 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, 2015; Chaudhuri et al., 2012; Lei and Rinaldo, 2015). Algorithm 1 is general enough to cover degree-corrected block model. We only need to modify the parameter estimation step. Similarly, we can use many different methods to estimate (g, B, ψ), such as spectral clustering, maximum likelihood, and least squares. Here we describe a rectangular spherical spectral clustering method.

Rectangular spherical spectral clustering:

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

1. Let $\widehat{U}$ be the $n \times \widetilde{K}$ matrix consisting of the top $\widetilde{K}$ right singular vectors of $A^{(1)}$.
2. Let $\widetilde{U}$ be the matrix obtained by scaling each row of $\widehat{U}$ to unit norm.

3. Output $\widehat{g}$ by applying the k-median clustering algorithm with $\widetilde{K}$ clusters to the rows of $\widetilde{U}$.

This algorithm is adapted from the spherical spectral clustering for full adjacency matrix in Lei and Rinaldo (2015). 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 4 below, the community information is contained in the normalized matrix $\widetilde{U}$, whereas the node activeness information is contained in the row norms of $\widehat{U}$.

The community recovery is obtained by solving a k-median clustering problem

$$\min_{u_1,\ldots,u_{\widetilde{K}}\in\mathbb{R}^{\widetilde{K}},\ g\in\{1,\ldots,\widetilde{K}\}^n} \sum_{i=1}^{n} \|\widetilde{u}_i - u_{g_i}\|, \tag{7}$$

where $\widetilde{u}_i$ is the ith row of $\widetilde{U}$. Let $(\widehat{u}_1,\ldots\widehat{u}_K,\widehat{g})$ be a solution to (7), then the estimated community is $\widehat{g}$. Finding the global optimum of (7) is computationally hard. But our method and corresponding theory also work for any approximate solution whose object value is within a constant factor from the global optimum. Such approximate solutions can be found using efficient algorithms (Charikar et al., 1999; Li and Svensson, 2013). If the matrix $\widehat{U}$ 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 $\widehat{U}$ is negligible under mild conditions.

Now we define a community-normalized version of ψ

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

Recall that ψ and B are only identifiable up to a scaling factor, for our purpose of estimating P_{ij}, it is sufficient to obtain estimates of ψ' and B', the correspondingly scaled version of B.

We propose to estimate ψ'_i as the ℓ_2 norm of the ith row of $\widehat{U}$:

$$\widehat{\psi}'_i = \left(\sum_{j=1}^{\widetilde{K}} \widehat{U}_{ij}^2\right)^{1/2}. \tag{8}$$

We will show, in the proof of Theorem 4 below, that $\widehat{\psi}'$ is a good estimate of ψ' under appropriate conditions. We estimate B' with the plug-in estimator:

$$\widehat{B}'_{k,k'} = \begin{cases} \dfrac{\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} A_{ij}}{\sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}} \widehat{\psi}'_i\widehat{\psi}'_j}\,, & k\neq k'\,, \\ \dfrac{\sum_{i,j\in\mathcal{N}_{1,k}, i<j} A_{ij} + \sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{2,k}} A_{ij}}{\sum_{i,j\in\mathcal{N}_{1,k}, i<j} \widehat{\psi}'_i\widehat{\psi}'_j + \sum_{i\in\mathcal{N}_{1,k}, j\in\mathcal{N}_{2,k}} \widehat{\psi}'_i\widehat{\psi}'_j}\,, & k=k'\,. \end{cases} \tag{9}$$

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

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

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

(A3) $\inf_{1\le i\le n}\psi_i \ge \psi_0$ for a positive constant ψ_0.

Theorem 4. *Under (A1)–(A3), assume B_0 is non-singular and $(\widehat{g}, \widehat{B}, \widehat{\psi})$ is obtained by spherical spectral clustering combined with Equations (8) and (9), then for each fold split we have*

(a) if $\rho_n \ge c\log n/n$ for a large enough constant c, then, with probability tending to one, $\widehat{g}$ agrees with g on all but $O(\sqrt{n/\rho_n})$ nodes;

(b) if $\rho_n = \omega(n^{-1/3})$, then $\widehat{P}_{ij} = P_{ij}(1+o_p(1))$ for all but a vanishing proportion of node pairs.

Theorem 4 is proved in Appendix A.2. Part (a) establishes approximate consistency of spherical spectral clustering applied on the rectangle fitting set of node pairs. Part (b) requires a larger average edge probability so that the estimation error of $\widehat{B}$ is well-controlled. In this case, part (a) of the theorem suggests that the proportion of mis-clustered nodes is $o(n^{-1/3})$.

4.2 Choosing model type and K simultaneously

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. The problem of choosing between the regular stochastic block model and the degree corrected model is first considered in Yan et al. (2014) with a focus on Poisson networks. They use a likelihood ratio approach combined with belief propagation for fast computation. In the context of binary network models, our V-fold NCV

can simultaneously choose between the regular stochastic block model and the degree corrected block model, and determine the number of communities. To this end, one just needs to calculate the regular stochastic block model validation error $\widehat{L}_{\text{sbm}}(A, \widetilde{K})$, and the degree corrected block model validation error $\widehat{L}_{\text{dcbm}}(A, \widetilde{K})$, for a collection of values of $\widetilde{K}$ as described in Algorithm 1. 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 5.

5. NUMERICAL EXPERIMENTS

5.1 Simulation studies

We first demonstrate the performance of NCV in both stochastic block models (Simulation 1) and degree corrected block models (Simulation 2), then we compare NCV with other state-of-art methods (Simulation 3).

Simulation 1: edge sparsity and community imbalance. This simulation is designed to investigate the performance of NCV for stochastic block models under different levels of edge sparsity and community size imbalance. We set 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 levels are chosen at $r \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, so that for $n = 1000$ the smallest expected degree ranges from 12 to 400. For community sizes, we set the size of the smallest community to be n_1, and the size of each of the remaining $K - 1$ communities to be $(n - n_1)/(K - 1)$. We generate edges according to the stochastic block model (1). For each combination of (r, K, n_1), three-fold NCV model selection based on simple spectral clustering is carried out for 200 independently drawn adjacency matrices. Figure 1 shows the proportion of correct model selection among these 200 repetitions as functions of r for different n_1 and $K \in \{2, 3, 4\}$. As expected, the performance is better as r and n_1 increase. In particular, for $K = 2$, in the most balanced case ($n_1 = 500$), NCV can perfectly choose the true number of clusters even in the sparsest case $r = 0.01$, whereas in the most imbalanced case ($n_1 = 100$), 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, NCV needs moderate expected node degrees for successful model selection. Community recovery for a given K is an integrated step in the proposed NCV method, so the

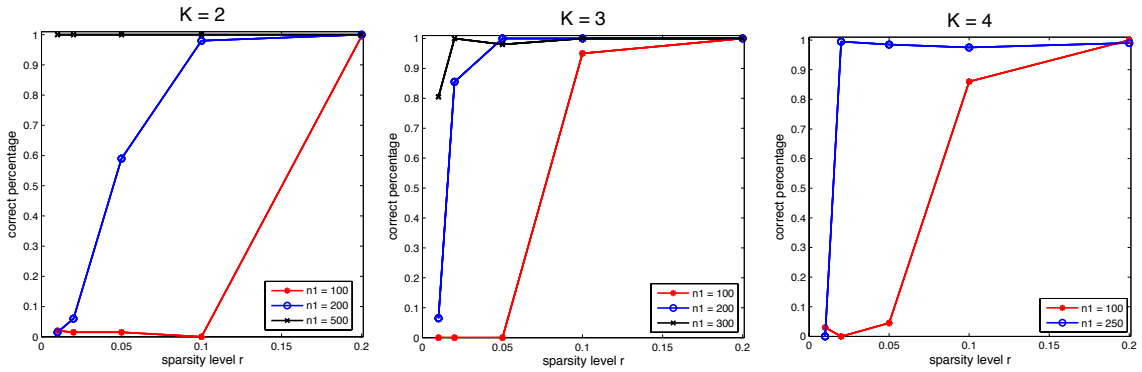

Figure 1: Results for **Simulation 1**: the proportion of correct estimate of K for stochastic block models over 200 repetitions, for $K = 2, 3, 4$, under sparsity levels $r \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, and various sizes of the first community n_1. The number of nodes is 1000.

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: degree corrected block models. This simulation is designed to demonstrate the performance of NCV in selecting between the stochastic block model and the degree-corrected block model with simultaneous selection of K. We use B matrices whose diagonal is 0.25 and off-diagonal is 0.1, which give a moderate sparsity level for stochastic block models. For degree-corrected block model, the degree parameter ψ is generated from Uniform$(0.2, 1)$, and normalized to have block-wise maximum value of 1. The network is much sparser due to the degree parameter ψ. We use three-fold NCV based on spectral clustering 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 $\widehat{T} = T$ and proportion of correct choice of K given correct model type selection. We generate 200 independent adjacency matrices from both the stochastic block model and the degree corrected block model, for each combination of $K = 1, 2, 3, 4$ and $n = 300, 600, 1200$. When the true model type is the stochastic block model, NCV can almost always 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 true model is the 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. With a smaller n, the method sometimes fails to select

Table 1: Results for Simulation 2: proportion of selecting the correct model type, and choosing the correct K given correct model type selection, over 200 independent simulations. The true models are generated from stochastic block models (SBM) or degree corrected block models (DCBM), for $K = 1, 2, 3, 4$ and $n = 300, 600, 1200$.

		SBM				*DCBM*			
		$K=1$	2	3	4	$K=1$	2	3	4
$n=300$	$\widehat{P}(\widehat{T}=T)$	1	1	1	1	1	0.825	0.605	0.450
	$\widehat{P}(\widehat{K}=K\|\widehat{T}=T)$	0.99	1	0.995	0.905	0.995	0.424	0	0
$n=600$	$\widehat{P}(\widehat{T}=T)$	1	1	1	1	1	1	0.99	0.99
	$\widehat{P}(\widehat{K}=K\|\widehat{T}=T)$	1	1	1	0.995	1	1	0.374	0
$n=1200$	$\widehat{P}(\widehat{T}=T)$	1	1	1	1	1	1	1	1
	$\widehat{P}(\widehat{K}=K\|\widehat{T}=T)$	0.99	1	0.995	0.985	1	1	1	0.985

the right K even when the correct DCBM model is selected. For the case of $n = 300, K = 2$, when the true model $DCBM$ is selected, the distribution of the selected K is 57% for $K = 1$, 43% for $K = 2$, and 1% for $K = 5$. For the case of $(n = 600, K = 3)$, when the true model $DCBM$ is selected, the distribution of the selected K is 62% for $K = 1$, 37% for $K = 2$, and 1% for $K = 4$.

Simulation 3: general block-wise edge probability structures and comparison to other methods. This simulation is designed to further investigate the overall performance of NCV under general stochastic block models, and meanwhile to compare it with the conditional BIC method proposed in Handcock et al. (2007); Airoldi et al. (2008), the variational Bayes method (Latouche et al., 2012), and the recursive testing procedure proposed in Bickel and Sarkar (2016). We consider four elements in designing the simulation.

1. Network size: $n \in \{600, 1200\}$, where n is the number of nodes in the network.

2. Network sparsity: $r \in \{0.05, 0.1, 0.15, 0.2\}$, where r controls the overall network sparsity.

3. Number of communities: $K \in \{2, 3, 4, 5\}$.

4. Type of community-wise edge probability: assortative-mixing and general structure. In the assortative-mixing model, the within-community edge probability is higher than between-

community edge probability. In particular, we set $B_{k,k'} = 2r$ for $k \neq k'$ and $B_{k,k} = 3r$. In the general structure model, we also set $B_{k,k'} = 2r$ but the diagonal entries are $(3,1) \times r$ for $K = 2$, $(3,2,1) \times r$ for $K = 3$, $(3,3,1,1) \times r$ for $K = 4$, and $(3,3,2,1,1) \times r$ for $K = 5$.

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 apply four different versions of three-fold NCV. The first one is based on spectral clustering, called `ncv.spec`. The second one, called `ncv2.spec`, repeats `ncv.spec` 20 times and chooses the most frequent output. The third one, called `ncv.ls`, is NCV combined with community recovery using the least squares estimate. We implement the least squares method using the label-switching algorithm as described in Bickel and Chen (2009); Stephens (2000). The fourth one, called `ncv2.ls`, repeats `ncv.ls` 20 times and chooses the most frequent output. For comparison, we implement the conditional BIC method first proposed by Handcock et al. (2007), where the membership is recovered using spectral clustering (`bic.spec`) or the least squares method (`bic.ls`). This method combines a conditional log-likelihood with a BIC-type penalty. We also compare with two other methods, the variational Bayes method (`vB`) developed by Latouche et al. (2012), and the recursive bipartition algorithm developed in Bickel and Sarkar (2016) with type I error lavel $\alpha = 0.01$ (`BiPart`). The recursive bipartition method divides 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 $K = 1$.

In Tables 2 and 3, we present the proportion of correct model selection over 200 independently generated networks for each combination of simulation setting (Due to space limitation, we omitted the results for $K = 2$ where all the methods work well, and the results for the recursive bipartition algorithm (`BiPart`), which is overall less competitive). As expected, all methods benefit from a larger sample size or a denser network. The task of choosing K gets harder as the true number of communities gets larger.

Under the assortative mixing model (Table 2), the B matrices are well-conditioned so that both least squares and spectral methods can successfully recover the hidden community. The proposed NCV methods with repetitions work best. The two conditional BIC methods are on a par with NCV with repetitions. The single round NCV methods are less stable when $n = 600$, but still work well when $n = 1200$. The variational Bayes method tends to require a denser network in order to

perform well. For example, when $K = 5$, it started to work when $r = 0.15$ and $n = 1200$.

Under the general structure model (Table 3), the least squares based NCV with repetitions (`ncv2.ls`) has the most satisfactory performance. The least squares based conditional BIC method is very close. The variational Bayes method still exhibits a stronger requirement on network density, when compared to the two leading methods. In this setting spectral based methods, such as `ncv2.spec`, `ncv.spec`, and `bic.spec`, are not as good as their least squares based counterparts. This is because of a much worse condition number of the B matrix, which makes spectral clustering less accurate. This observation reflects the theoretical superiority of likelihood and least squares methods in community recovery (Amini et al., 2013; Abbe et al., 2016; Gao et al., 2015).

Overall, NCV has very satisfactory performance for selecting K in SBM, under different structures of B. we found that for a relatively easy B with reasonable dense data or large n, the accuracy of one split ncv is almost close to 1 and so as multiple splits. In the case of relatively difficult B and smaller n, we do see some variability in one split of ncv, where the multiple splits helps to largely reduce the variability and improve the accuracy to almost 1. For example, the setting in with general B structure, $K = 4$, $r = 0.1$ and $n = 600$, the selected K is highly concentrated on the true $K = 4$, with a small proportion of over selection of $K = 5$. The conditional BIC method has comparable performance. However, as we have seen in *Simulation 2*, NCV is directly applicable for general model comparison purposes, while the conditional BIC method cannot be used to compare two different models such as SBM and DCBM. If condition on the estimated ψ, then the conditional log-likelihoods of SBM and DCBM are not comparable, as DCBM is always preferred. How to appropriately incorporate the conditional perspective into BIC-type penalties and calibrate the measure of model complexity remains an open problem.

5.2 Data examples

Here we apply the NCV method to two well-known network data sets with community structure.

Data example 1: political weblogs. The political blog data was collected and analyzed in Adamic and 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 with two communities is believed to fit

Table 2: Results for Simulation 3, Assortative Mixing. Reported are proportions of correct model selection in 200 independently generated data. Community sizes are equal, with $B(k,k') = 2r$ for $k \neq k'$ and $B(k,k) = 3r$.

n	K	r	*ncv.ls*	*ncv*2.*ls*	*ncv.spec*	*ncv*2.*spec*	*vB*	*bic.ls*	*bic.spec*
600	3	0.05	0.00	0.00	0.01	0.00	0.00	0.00	0.00
		0.1	0.95	1.00	1.00	1.00	0.92	1.00	1.00
		0.15	0.97	1.00	1.00	1.00	1.00	1.00	1.00
		0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.1	0.23	0.29	0.55	0.60	0.00	0.46	0.04
		0.15	0.96	1.00	1.00	1.00	0.99	1.00	1.00
		0.2	0.97	1.00	0.99	1.00	1.00	1.00	1.00
	5	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.15	0.78	1.00	0.78	0.96	0.06	1.00	0.95
		0.2	0.94	1.00	1.00	1.00	1.00	1.00	1.00
1200	3	0.05	0.92	1.00	0.99	1.00	0.02	0.99	1.00
		0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.05	0.01	0.00	0.15	0.03	0.00	0.03	0.00
		0.1	0.99	1.00	1.00	1.00	0.98	1.00	1.00
		0.15	0.99	1.00	1.00	1.00	1.00	1.00	1.00
		0.2	1.00	1.00	0.98	1.00	1.00	1.00	1.00
	5	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.1	0.95	1.00	1.00	1.00	0.02	1.00	1.00
		0.15	0.95	1.00	1.00	1.00	1.00	1.00	1.00
		0.2	1.00	1.00	0.96	1.00	1.00	1.00	1.00

Table 3: Results for Simulation 3, General Structure. Reported are proportion of correct model selection in 200 independently generated data, with equal community sizes. The B matrix is chosen such that $B(k,k') = 2r$ for $k \neq k'$, and the diagonal entries are $(3,1) \times r$ for $K = 2$, $(3,2,1) \times r$ for $K = 3$, $(3,3,1,1) \times r$ for $K = 4$, and $(3,3,2,1,1) \times r$ for $K = 5$.

n	K	r	*ncv.ls*	*ncv*2.*ls*	*ncv.spec*	*ncv*2.*spec*	*vB*	*bic.ls*	*bic.spec*
600	3	0.05	0.15	0.00	0.00	0.00	0.00	0.04	0.00
		0.1	0.84	1.00	0.08	0.02	1.00	1.00	0.08
		0.15	0.96	1.00	0.80	1.00	1.00	1.00	0.98
		0.2	0.98	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.1	0.70	0.99	0.04	0.00	0.33	0.96	0.00
		0.15	0.97	1.00	0.34	0.08	1.00	1.00	0.32
		0.2	0.97	1.00	0.91	1.00	1.00	1.00	0.99
	5	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.15	0.82	1.00	0.05	0.00	0.64	1.00	0.00
		0.2	0.94	1.00	0.18	0.08	1.00	1.00	0.15
1200	3	0.05	0.64	0.94	0.03	0.00	1.00	1.00	0.00
		0.1	0.98	1.00	0.96	1.00	1.00	1.00	1.00
		0.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.05	0.40	0.52	0.00	0.00	0.03	0.68	0.00
		0.1	0.88	1.00	0.70	0.88	1.00	0.86	0.90
		0.15	0.92	1.00	1.00	1.00	1.00	0.96	1.00
		0.2	0.98	1.00	1.00	1.00	1.00	1.00	1.00
	5	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.1	0.83	1.00	0.16	0.02	0.82	1.00	0.01
		0.15	0.99	1.00	0.42	0.42	1.00	1.00	0.34
		0.2	1.00	1.00	0.78	0.98	1.00	1.00	0.42

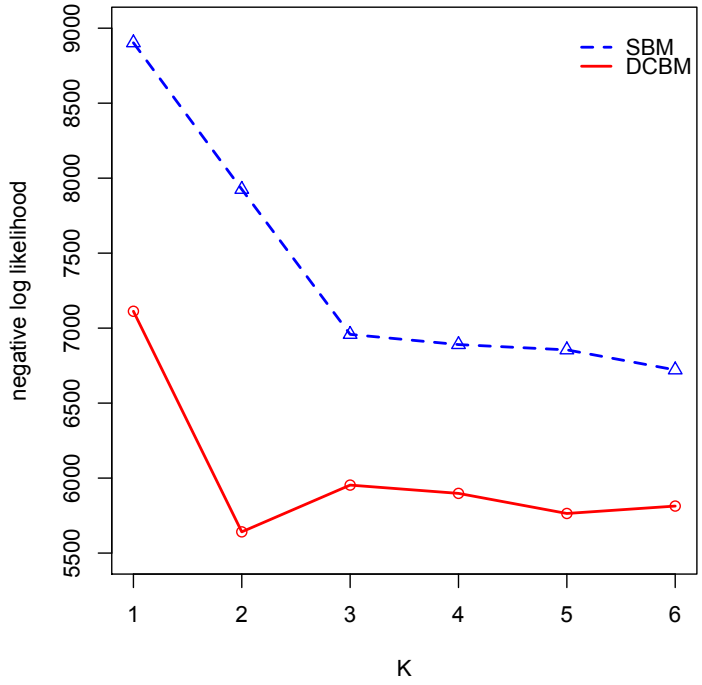

Figure 2: Results for **the political blogs data**: reporting the three-fold cross-validated negative log-likelihood of all candidate models from one random block splitting. Dashed line: stochastic block models; solid line: degree corrected block models. The results are consistent over 100 repeated random block splittings.

better than the stochastic block model (Karrer and Newman, 2011; Zhao et al., 2012; Jin, 2015). 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 cross-validated negative log-likelihood for all candidate models is plotted in Figure 2 for a typical block splitting. Despite the randomness in block splitting, the NCV is very stable when applied on this data set. We repeated the NCV procedure 100 times using independent random block splittings. The NCV selected DCBM with $K = 2$ in 99 out of 100 repetitions, where the one failure was due to non-convergence of k-means in spectral clustering.

Data example 2: political books. We consider a co-purchasing network data of political books sold on `Amazon.com`. The data was collected by Krebs (2004). It has 105 nodes, each representing a political book. An edge between a pair of nodes indicates frequent co-purchasing of these two

books (defined as the "customers who bought this item also bought..." feature of `Amazon.com`). The average degree is 4.2. The nodes are manually labeled by Mark Newman in one of the three categories: "neutral" (13 nodes), "liberal" (43 nodes), "conservative" (49 nodes). This is a rather small data set, and it would be unrealistic to estimate a DCBM with three communities from such a small sample. So we focus on fitting this data set using SBM's. We repeated the three-fold NCV 100 times using independent random block splittings, the most frequent output is $\widehat{K} = 3$. We also tried the most frequent output from 20 random splittings as used in Simulation 2, combined with three-, four-, and five-fold NCV. They all gave the same estimate $\widehat{K} = 3$. In comparison, the conditional BIC approach chooses $\widehat{K} = 4$. When spectral clustering is used for community recovery with $K = 3$, the result evenly partitions the nodes into three groups, where two of the groups contain mostly conservative or liberal nodes, and the other group contains a mixture of 11 neutral nodes, 10 liberal nodes, and 16 conservative nodes. The interpretation is that while extreme nodes are easy to cluster, the less extreme ones tend to mix with the neutral ones. A similar phenomenon on the same data set has been observed and reported in Newman (2006).

6. DISCUSSION

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 block model are good examples that satisfy 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 and 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 to allow for estimation using only a subset of rows in A.

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 a 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 ratio between the sizes of fitting and testing samples 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 fitting stage but will 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 fitting and testing sample size ratio.

A. PROOFS

A.1 Proof of Theorem 2

Proof. We focus on a single fold in NCV. The claimed results follow by summing over all folds. Again let $\mathcal{N}_1$ be the nodes corresponding to the fitting part and $\mathcal{N}_2$ be those in the testing part. Let $I_l = \{i \in \mathcal{N}_2 : g_i = l,\}$ $(1 \le l \le K)$ and $\widehat{I}_k = \{i \in \mathcal{N}_2 : \widehat{g}_i = k\}$ $(1 \le k \le \widetilde{K})$.

Case 1: $\widetilde{K} < K$.

According to tail probability inequalities for hypergeometric distributions (see Lemma 5), with overwhelming probability over the random data splitting, the testing block $A^{(vv)}$ is also a realization of a K-block SBM satisfying assumptions A1–A2, with a different constant π_0'.

When $\widetilde{K} < K$, there exist $1 \le k \le \widetilde{K}$, $1 \le l_1 < l_2 \le K$ such that $|\widehat{I}_k \cap I_{l_j}| \ge |I_{l_j}|/\widetilde{K} \ge \pi_0' n/|\widetilde{K}|$ for $j = 1, 2$. Because B_0 does not have identical rows, there exists $1 \le l_3 \le K$ such that $B_0(l_1, l_3) \ne B_0(l_2, l_3)$. There exists a k' such that $|\widehat{I}_{k'} \cap I_{l_3}| \ge |I_{l_3}|/\widetilde{K} \ge \pi_0' n/\widetilde{K}$. We focus on the case $k \ne k'$ and l_1, l_2, l_3 are distinct. The other cases, such as $k = k'$ or $l_1 = l_3$, or both, can be dealt with similarly. Without loss of generality, assume $k = 1, k' = 2$, $l_1 = 1, l_2 = 2, l_3 = 3$.

Let $\mathcal{T}_{k,k',l,l'}$ be the set of unique pairs (i, j) such that $i \in \widehat{I}_k \cap I_l$, $j \in \widehat{I}_{k'} \cap I_l$, and $i \ne j$ (that is, (i, j) and (j, i) shall be counted as one unique pair in case of $k = k'$ and $l = l'$). Let $\widehat{p}$ be the average of A_{ij} over $(i, j) \in \mathcal{T}_{1,2,1,3} \cup \mathcal{T}_{1,2,2,3}$ and $\widehat{p}_{k,k',l,l'}$ be the average of A_{ij} over $(i, j) \in \mathcal{T}_{k,k',l,l'}$. Then, letting $L^{(v)}(A)$ be the terms in $L(A)$ corresponding to the vth fold,

$$\begin{aligned}
&\widehat{L}^{(v)}(A, \widetilde{K}) - L^{(v)}(A) \\
&= \sum_{k,k',l,l'} \sum_{(i,j) \in \mathcal{T}_{k,k',l,l'}} \left[(A_{ij} - \widehat{P}_{ij})^2 - (A_{ij} - P_{ij})^2 \right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k,k',l,l'} \sum_{(i,j)\in\mathcal{T}_{k,k',l,l'}} \left[(A_{ij}-\widehat{B}_{k,k'})^2-(A_{ij}-B_{l,l'})^2\right] \\
&= \sum_{(i,j)\in\mathcal{T}_{1,2,1,3}} \left[(A_{ij}-\widehat{B}_{1,2})^2-(A_{ij}-B_{1,3})^2\right] + \sum_{(i,j)\in\mathcal{T}_{1,2,2,3}} \left[(A_{ij}-\widehat{B}_{1,2})^2-(A_{ij}-B_{2,3})^2\right] \\
&\quad + \sum_{(k,k',l,l')\notin\{(1,2,1,3),(1,2,2,3)\}} \sum_{(i,j)\in\mathcal{T}_{k,k',l,l'}} \left[(A_{ij}-\widehat{B}_{k,k'})^2-(A_{ij}-B_{l,l'})^2\right] \\
&\geq \sum_{(i,j)\in\mathcal{T}_{1,2,1,3}} \left[(A_{ij}-\widehat{p})^2-(A_{ij}-B_{1,3})^2\right] + \sum_{(i,j)\in\mathcal{T}_{1,2,2,3}} \left[(A_{ij}-\widehat{p})^2-(A_{ij}-B_{2,3})^2\right] \\
&\quad + \sum_{(k,k',l,l')\notin\{(1,2,1,3),(1,2,2,3)\}} \sum_{(i,j)\in\mathcal{T}_{k,k',l,l'}} \left[(A_{ij}-\widehat{p}_{k,k',l,l'})^2-(A_{ij}-B_{l,l'})^2\right] \\
&= I + II + III.
\end{aligned}
$$

Let $\lambda = |\mathcal{T}_{1,2,1,3}|/(|\mathcal{T}_{1,2,1,3}+\mathcal{T}_{1,2,2,3}|)$. Then $\pi_0'^2/(\widetilde{K}^2+\pi_0'^2) \leq \lambda \leq \widetilde{K}^2/(\widetilde{K}^2+\pi_0'^2)$, and $\widehat{p} = \lambda\widehat{p}_{1,2,1,3}+(1-\lambda)\widehat{p}_{1,2,2,3}$. If $|\mathcal{T}_{k,k',l,l'}|>0$, using Bernstein's inequality we have

$$
|\widehat{p}_{k,k',l,l'}-B_{l,l'}| = O_P\left(\sqrt{\frac{\rho_n}{|\mathcal{T}_{k,k',l,l'}|}}\right). \tag{A.1}
$$

$$
\begin{aligned}
I =& |\mathcal{T}_{1,2,1,3}|\left[(\widehat{p}-\widehat{p}_{1,2,1,3})^2-(\widehat{p}_{1,2,1,3}-B_{1,3})^2\right] \\
=& |\mathcal{T}_{1,2,1,3}|\left[(1-\lambda)^2(\widehat{p}_{1,2,1,3}-\widehat{p}_{1,2,2,3})^2-(\widehat{p}_{1,2,1,3}-B_{1,3})^2\right] \\
\geq& |\mathcal{T}_{1,2,1,3}|\Bigg[\frac{(1-\lambda)^2}{2}(B_{1,3}-B_{2,3})^2-(1-\lambda)^2\left((\widehat{p}_{1,2,1,3}-B_{1,3})-(\widehat{p}_{1,2,2,3}-B_{2,3})\right)^2 \\
&\qquad -(\widehat{p}_{1,2,1,3}-B_{1,3})^2\Bigg] \\
\geq& |\mathcal{T}_{1,2,1,3}|\Bigg[\frac{(1-\lambda)^2}{2}(B_{1,3}-B_{2,3})^2-2(1-\lambda)^2\left(\widehat{p}_{1,2,1,3}-B_{1,3}\right)^2 \\
&\qquad -2(1-\lambda)^2\left(\widehat{p}_{1,2,2,3}-B_{2,3}\right)^2-(\widehat{p}_{1,2,1,3}-B_{1,3})^2\Bigg] \\
\geq& c(n\rho_n)^2+O_P(\rho_n).
\end{aligned}
$$

where the constant c depending on π_0, B_0 and $\widetilde{K}$ only, the first term in the last inequality comes from the fact that $|\mathcal{T}_{1,2,1,3}| \geq \pi_0^2 n^2/\widetilde{K}^2$ and $|B_{1,2}-B_{1,3}| \geq c'\rho_n$ for some c' depending only on B_0, the second term comes from (A.1) and the fact that $|T_{1,2,1,3}| \asymp |T_{1,2,2,3}|$.

Similarly,

$$
II \geq c(n\rho_n)^2+O_P(\rho_n).
$$

and

$$III \geq - \sum_{(k,k',l,l') \notin \{(1,2,1,3),(1,2,2,3)\}} |\mathcal{T}_{k,k',l,l'}| (\widehat{p}_{k,k',l,l'} - B_{l,l'})^2 = O_P(\rho_n)\,.$$

Case 2: $\widetilde{K} = K$ with exactly consistent recovery

In this case we focus on the event $\widehat{g} = g$, which has $1 - o(1)$ probability. Then $\epsilon_n := \sup_{1 \leq l,l' \leq K} |\widehat{B}_{l,l'} - B_{l,l'}| = O_P(\sqrt{\rho_n}/n)$ by Bernstein's inequality and $\widehat{B}_{l,l'}$ are independent with A_{ij} for (i,j) in the testing set.

For $1 \leq k, k' \leq K$, let $\mathcal{T}_{l,l'}$ be the collection of pairs (i,j) in the testing set such that $i \in I_k$, $j \in I_{k'}$ and $i \neq j$.

$$\begin{aligned}
\left|\widehat{L}^{(v)}(A, \widetilde{K}) - L^{(v)}(A)\right| \leq & \sum_{k,k'} \sum_{(i,j) \in \mathcal{T}_{k,k'}} \left|(A_{ij} - \widehat{B}_{k,k'})^2 - (A_{ij} - B_{k,k'})^2\right| \\
= & \sum_{k,k'} \sum_{(i,j) \in \mathcal{T}_{k,k'}} \left|-2A_{ij}(\widehat{B}_{k,k'} - B_{k,k'}) + (\widehat{B}_{k,k'} + B_{k,k'})(\widehat{B}_{k,k'} - B_{k,k'})\right| \\
\leq & \sum_{k,k'} \sum_{(i,j) \in \mathcal{T}_{k,k'}} [2\epsilon_n A_{ij} + (2\rho_n + \epsilon_n)\epsilon_n] \\
\leq & O_p(n^2\rho_n)\epsilon_n + n^2(2\rho_n + \epsilon_n)\epsilon_n = O_P(n\rho_n^{3/2})\,.
\end{aligned}$$

Because $n\rho_n^{3/2} = o((n\rho_n)^2)$, we have

$$\begin{aligned}
P(\widehat{K} < K) \leq & \sum_{\widetilde{K} < K} P\left[\widehat{L}(A, \widetilde{K}) < \widehat{L}(A, K)\right] \\
\leq & \sum_{\widetilde{K} < K} \sum_{1 \leq v \leq V} P\left[\widehat{L}^{(v)}(A, \widetilde{K}) < \widehat{L}^{(v)}(A, K)\right] = o(1)\,.
\end{aligned}$$

This proves part (a) of Corollary 3.

Case 3: $\widetilde{K} = K$ with approximately consistent recovery

Without loss of generality we assume the optimal permutation to match $\widehat{g}$ and g is the identity. Let $\widehat{\mathcal{U}}_{k,k'}$ be the set of distinct pairs (i,j) in the fitting subset such that $\widehat{g}_i = k$, $\widehat{g}_i = k'$, and $i \neq j$. Let $\mathcal{U}_{k,k'}$ be the set of distinct pairs (i,j) in the fitting subset such that $g_i = k$, $g_i = k'$, and $i \neq j$. Let $S = \bigcup_{k=1}^K \widehat{I}_k \Delta I_k$, $\mathcal{S} = \bigcup_{k,k'} \widehat{\mathcal{U}}_{k,k'} \Delta \mathcal{U}_{k,k'}$. On the event $|S| \leq \eta_n n$, under Assumptions (A1) and (A2) we have $|\mathcal{S}| = O(\eta_n n^2)$, and hence $|\widehat{\mathcal{U}}_{k,k'}| = |\mathcal{U}_{k,k'}|(1 + O(\eta_n))$ for all k, k'.

$$\left|\widehat{B}_{k,k'} - B_{k,k'}\right|$$

$$
\begin{aligned}
=&\left|\frac{1}{|\widehat{\mathcal{U}}_k|}\left(\sum_{(i,j)\in\mathcal{U}_{k,k'}} A_{ij}\right) - B_{k,k'}\right| + \frac{1}{|\widehat{\mathcal{U}}_{k,k'}|}\sum_{(i,j)\in\widehat{\mathcal{U}}_{k,k'}\Delta\mathcal{U}_{k,k'}} A_{ij} \\
\leq&\frac{|\mathcal{U}_{k,k'}|}{|\widehat{\mathcal{U}}_{k,k'}|}\left|\frac{1}{|\mathcal{U}_{k,k'}|}\sum_{(i,j)\in\mathcal{U}_{k,k'}} A_{ij} - B_{k,k'}\right| + \left|\frac{|\mathcal{U}_{k,k'}|}{|\widehat{\mathcal{U}}_{k,k'}|} - 1\right| B_{k,k'} + \frac{1}{|\widehat{\mathcal{U}}_{k,k'}|}\sum_{(i,j)\in\widehat{\mathcal{U}}_{k,k'}\Delta\mathcal{U}_{k,k'}} A_{ij} \\
\leq&(1+O(\eta_n))O_P\left(\sqrt{B_{k,k'}/|\mathcal{U}_{k,k'}|}\right) + O(\eta_n)B_{k,k'} + (1+O(\eta_n))\eta_n \\
=&O_P(\rho_n^{1/2}n^{-1} + \eta_n)\,.
\end{aligned}
$$

Focusing on the vth fold in cross-validation, we similarly have, letting $\epsilon_n = \sup_{k,k'}|\widehat{B}_{k,k'} - B_{k,k'}|$,

$$
\left|\widehat{L}^{(v)}(A,\widetilde{K}) - L^{(v)}(A)\right| \leq O_p(n^2(\rho_n+\epsilon_n)\epsilon_n) = O_p(n\rho_n^{3/2} + n^2\rho_n\eta_n)\,. \qquad \blacksquare
$$

A.2 Proof of Theorem 1 and Theorem 4

Next we prove Theorem 1 and Theorem 4. Let P be the $n\times n$ matrix such that $P_{ij} = B_{g_ig_j}$. For a particular fold split in NCV, let $A^{(1)}$ and $P^{(1)}$ be the testing rectangular submatrix of A and P respectively. 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.

Preliminary results The following technical lemmas are needed in the proof of Theorem 1 and Theorem 4.

Lemma 5 (Size of split community)**.** *Under Assumption (A1), for n large enough we have $\min_k n_{1,k}^* \geq \pi_0 n/(2V)$, 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 6 (Spectral norm error of partial adjacency matrix)**.** *Let A be the adjacency matrix generated from a degree corrected block model satisfying Assumption A2, with $\rho_n \geq c\log n/n$ for a positive constant c. Let $\widetilde{A}$ be an arbitrary subset of rows of A and $\widetilde{P}$ be the corresponding submatrix*

of P. We have, for some constant C,

$$P\left(\|\widetilde{A}-\widetilde{P}\| \le C\sqrt{n\rho_n}\right) \ge 1-n^{-1}/2\,.$$

Proof. Observe that $\|\widetilde{A}-\widetilde{P}\| \le \|A-P\|$. The claimed result follows easily from Theorem 5.2 of Lei and Rinaldo (2015), where it has been shown that with high probability $\|A-P\| \le C\sqrt{n\rho_n}$. ∎

Lemma 6 covers both regular SBM and DCBM. It implies that $\|A^{(1)}-P^{(1)}\| \le C\sqrt{n\rho_n}$ with high probability.

Lemma 7 (Singular subspace error bound)**.** *Let $\widehat{M}$, M be two matrices of same dimension, and $\widehat{U}$ and U be $n\times K$ orthonormal matrices corresponding to the top K right singular vectors of $\widehat{M}$ and M, respectively. Then there exists a $K\times K$ orthogonal matrix Q such that*

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

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

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.

Lemma 8 (Lemma 5.3 of Lei and Rinaldo (2015))**.** *Let $\widehat{U}$ 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 $\widehat{g}$ be the output of a k-means clustering algorithm on $\widehat{U}$, with objective value no larger than a constant factor of the global optimum. Assume that $\|\widehat{U}-U\|_F \le cn\delta$ for some small enough constant c. Then $\widehat{g}$ agrees with g on all but $c^{-1}\|\widehat{U}-U\|_F\delta^{-1}$ nodes after an appropriate label permutation.*

For the proof of Theorem 4, we need the analogous version of Lemma 8 for the k-median algorithm, which is a simple adaptation of Lemma 8. For a matrix M, $\|M\|_{2,1}$ denotes the sum of ℓ_2 norms of the rows in M.

Lemma 9. *Let $\widehat{U}$ 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 $\widehat{g}$ be the output of a k-median clustering algorithm on $\widehat{U}$, with objective value no larger than a constant factor of the global optimum. Assume that $\|\widehat{U} - U\|_{2,1} \leq cn\delta$ for some small enough constant c and that g satisfies Assumption A2. Then $\widehat{g}$ agrees with g on all but $c^{-1}\|\widehat{U} - U\|_{2,1}\delta^{-1}$ nodes after an appropriate label permutation.

Proof of Theorem 1

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 $P^{(1)} = G^{(1)}BG^T = G\widetilde{B}\widetilde{G}^T$, where $\widetilde{G}$ is an $n \times K$ matrix obtained by normalizing the columns of G. and $\widetilde{B}$ is a $K \times K$ matrix after corresponding column scaling of B. It is easy to check that $\widetilde{G}$ has orthonormal columns and hence the top K-dimensional right singular subspace of $P^{(1)}$ is spanned by $U = \widetilde{G}Q$, 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 $1/\sqrt{n}$.

We focus on the event that $\min_k n^*_{1,k} \geq \pi_0 n/(2V)$ and $\|A^{(1)} - P^{(1)}\| \leq C\sqrt{n\rho_n}$, which has probability at least $1 - n^{-1}$ according to Lemmas 5 and 6. Then it can be directly verified that $\sigma_K(P^{(1)}) \geq Cn\rho_n$ because the Kth largest singular values of both $G^{(1)}$ and G are of order at least $\sqrt{n}$. Then Lemma 7 implies that $\widehat{U}$ and U, the matrices consisting of $n \times K$ top singular vectors of $A^{(1)}$ and $P^{(1)}$, satisfy, with appropriate choice of Q,

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

Then applying Lemma 8, we know that the k-means clustering algorithm misclusters no more than C/ρ_n nodes. ∎

Proof of Theorem 4

Proof of Theorem 4. 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 $P^{(1)} = \Psi^{(1)}\widetilde{B}\Psi$, where $\widetilde{B}$ 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 $\widetilde{U}$ and $\widetilde{U}^*$ be the row-normalized versions of $\widehat{U}$

and U, respectively. Then $\widetilde{U}^*$ contains K distinct rows and the distance between any two distinct rows of $\widetilde{U}^*$ is $\sqrt{2}$.

We will focus on the event that $\min_k n^*_{1,k} \geq c_0\pi_0 n/2$ and $\|A^{(1)} - P^{(1)}\| \leq C\sqrt{n\rho_n}$, which has probability at least $1 - n^{-1}$ according to Lemmas 5 and 6. Applying Lemma 7 we have,

$$\|\widehat{U} - U\|_F^2 \leq C/(n\rho_n)\,,$$

where we take the matrix Q in Lemma 7 as identity. Because the minimum row norm of U is at least $\psi_0/\sqrt{n}$, the number of zero rows in $\widehat{U}$ is at most $\|\widehat{U} - U\|_F^2/(\psi_0/\sqrt{n})^2 = O(\rho_n^{-1}) = o(\sqrt{n/\rho_n})$. In the rest of the proof we can safely assume that $\widehat{U}$ has no zero rows.

Now let $u_i, \widehat{u}_i$ be the ith row of $U, \widehat{U}$, 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 A3,

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

Applying Lemma 9 to $\widetilde{U}$ and $\widetilde{U}^*$, we conclude that $\widehat{g}$ agrees with g on all but $O(\sqrt{n/\rho_n})$ nodes.

Recall that $\|u_i\| = \psi'_i$ for all i. Then Cauchy-Schwartz implies that

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

By Assumptions (A2) and (A3) we have $\inf_i \psi'_i \geq Cn^{-1/2}$ for some constant C.

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

$$|S_n^c| \leq \frac{\|\widehat{\psi}' - \psi'\|_1}{n^{-2/3}\rho_n^{-1/2}} \leq Cn^{2/3}\,. \tag{A.3}$$

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

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

It is obvious that $\widehat{P}^*_{ij} = \psi'_i\psi'_j\widehat{B}'^*_{g_ig_j} = (1 + o_P(1))\psi_i\psi_j B_{g_ig_j} = (1 + o_P(1))P_{ij}$. As a result, the claim in part (b) of the theorem follows if we can show that $\widehat{B}'_{k,k'} = (1 + o_P(1))\widehat{B}'^*_{k,k'}$ because for all but

a vanishing proportion of pairs (i,j) we have $(\widehat{g}_i,\widehat{g}_j)=(g_i,g_j)$ and $\widehat{\psi}'_i\widehat{\psi}'_j=\psi'_i\psi'_j(1+o(1))$ in view of (A.3). To this end, we compare

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

with

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

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 $n^{-1}\widehat{B}'^*_{k,k'}$ with

$$\sum_{i\in\mathcal{N}_{1,k},j\in\mathcal{N}_{1,k'}\cup\mathcal{N}_{2,k'}}(\sqrt{n}\psi'_i)(\sqrt{n}\psi'_j)\,. \tag{A.4}$$

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 (A.4) with the denominator of $n^{-1}\widehat{B}'_{k,k'}$. We have

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

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

REFERENCES

Abbe, E., Bandeira, A. S., and Hall, G. (2016), "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, 62, 471–487.

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

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed membership stochastic blockmodels," *The Journal of Machine Learning Research*, 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," *The Annals of Statistics*, 41, 2097–2122.

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. M. (2014), "A Tensor Approach to Learning Mixed Membership Community Models," *Journal of Machine Learning Research*, 15, 2239–2312.

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

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

Borgs, C., Chayes, J., and Smith, A. (2015), "Private graphon estimation for sparse graphs," in *Advances in Neural Information Processing Systems*, pp. 1369–1377.

Charikar, M., Guha, S., Tardos, É., and 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*, ACM, pp. 1–10.

Chatterjee, S. (2014), "Matrix estimation by universal singular value thresholding," *The Annals of Statistics*, 43, 177–214.

Chaudhuri, K., Chung, F., and 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., and Xu, H. (2012), "Clustering Sparse Graphs," in *Advances in Neural Information Processing Systems 25*, eds. Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K., pp. 2213–2221.

Daudin, J.-J., Picard, F., and Robin, S. (2008), "A mixture model for random graphs," *Statistics and computing*, 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," *Physical Review E*, 84, 066106.

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

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 Journal on Matrix Analysis and Applications*, 34, 23–39.

Gao, C., Lu, Y., Zhou, H. H., et al. (2015), "Rate-optimal graphon estimation," *The Annals of Statistics*, 43, 2624–2652.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 301–354.

Hoff, P. (2008), "Modeling homophily and stochastic equivalence in symmetric relational data," in *Advances in Neural Information Processing Systems*, pp. 657–664.

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

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

Jin, J. (2015), "Fast community detection by SCORE," *The Annals of Statistics*, 43, 57–89.

Josse, J. and Husson, F. (2012), "Selecting the number of components in principal component analysis using cross-validation approximations," *Computational Statistics & Data Analysis*, 56, 1869–1879.

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

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

Krebs, V. (2004), "Social Network Analysis software & services for organizations, communities, and their consultants," Http://www.orgnet.com/.

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

Latouche, P., Birmele, E., and Ambroise, C. (2012), "Variational Bayesian inference and complexity

control for stochastic block models," *Statistical Modelling*, 12, 93–115.

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

Lei, J. and Rinaldo, A. (2015), "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, 43, 215–237.

Lei, J. and Zhu, L. (2014), "A Generic Sample Splitting Approach for Refined Community Recovery in Stochastic Block Models," *arXiv preprint arXiv:1411.1469*.

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

Massoulié, L. (2014), "Community detection thresholds and the weak Ramanujan property," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, ACM, pp. 694–703.

McDaid, A. F., Murphy, T. B., Friel, N., and Hurley, N. J. (2013), "Improved Bayesian inference for the stochastic block model with application to large networks," *Computational Statistics & Data Analysis*, 60, 12–31.

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

Mossel, E., Neeman, J., and 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, 8577–8582.

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

Owen, A. B. and Perry, P. O. (2009), "Bi-cross-validation of the SVD and the nonnegative matrix factorization," *The Annals of Applied Statistics*, 564–594.

Peixoto, T. P. (2013), "Parsimonious module inference in large networks," *Physical review letters*, 110, 148701.

Rosvall, M. and Bergstrom, C. T. (2007), "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences*, 104,

7327–7331.

Saldana, D. F., Yu, Y., and Feng, Y. (2014), "How Many Communities Are There?" *arXiv preprint arXiv:1412.1684*.

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

Stephens, M. (2000), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 795–809.

Sussman, D. L., Tang, M., Fishkind, D. E., and Priebe, C. E. (2012), "A Consistent Adjacency Spectral Embedding for Stochastic Blockmodel Graphs," *Journal of the American Statistical Association*, 107, 1119–1128.

Sussman, D. L., Tang, M., and 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, 48–57.

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

Wang, Y. and Bickel, P. J. (2015), "Likelihood-based model selection for stochastic block models," *arXiv preprint arXiv:1502.02069*.

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

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

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

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

— (2012), "Consistency of community detection in networks under degree-corrected stochastic block models," *The Annals of Statistics*, 40, 2266–2292.