CONSISTENCY OF SPECTRAL CLUSTERING IN STOCHASTIC BLOCK MODELS

Author(s): Jing Lei and Alessandro Rinaldo

Source: *The Annals of Statistics*, February 2015, Vol. 43, No. 1 (February 2015), pp. 215-237

Published by: Institute of Mathematical Statistics

Stable URL: https://www.jstor.org/stable/43556513

The Annals of Statistics
2015, Vol. 43, No. 1, 215–237
DOI: 10.1214/14-AOS1274

CONSISTENCY OF SPECTRAL CLUSTERING IN STOCHASTIC BLOCK MODELS

BY JING LEI[1] AND ALESSANDRO RINALDO[2]

Carnegie Mellon University

We analyze the performance of spectral clustering for community extraction in stochastic block models. We show that, under mild conditions, spectral clustering applied to the adjacency matrix of the network can consistently recover hidden communities even when the order of the maximum expected degree is as small as $\log n$, with n the number of nodes. This result applies to some popular polynomial time spectral clustering algorithms and is further extended to degree corrected stochastic block models using a spherical k-median spectral clustering method. A key component of our analysis is a combinatorial bound on the spectrum of binary random matrices, which is sharper than the conventional matrix Bernstein inequality and may be of independent interest.

1. Introduction.

Network analysis is concerned with describing and modeling the joint occurrence of random interactions among actors in a given population of interest. In its simplest form, a network dataset over n actors is a simple undirected random graph on n nodes, where the edges encode the realized binary interactions among the nodes. Examples include social networks (friendship between Facebook users, blog following, twitter following, etc.), biological networks (gene network, gene-protein network), information network (email network, World Wide Web) and many others. A review of modeling and inference on network data can be found in Kolaczyk (2009), Newman (2010) and Goldenberg et al. (2010).

Among the many existing statistical models for network data, the stochastic block model, henceforth SBM, of Holland, Laskey and Leinhardt (1983) stands out for both its simplicity and expressive power. In a SBM, the nodes are partitioned into $K < n$ disjoint groups, or *communities*, according to some latent random mechanism. Conditionally on the realized but unobservable community assignments, the edges then occur independently with probabilities depending only on the community membership of the nodes, so that nodes from the same community will have higher average degree of connectivity among themselves than compared to the remaining nodes (see Section 2.1 for details). Because of its simple

Received September 2014.

[1] Supported by NSF Grant BCS-0941518, NSF Grant DMS-14-07771 and NIH Grant MH057881.

[2] Supported by AFOSR and DARPA Grant FA9550-12-1-0392 and NSF CAREER Grant DMS-11-49677.

MSC2010 subject classifications. 62F12.

Key words and phrases. Network data, stochastic block model, spectral clustering, sparsity.

analytic form and its ability to capture the emergence of communities, a feature commonly observed in real network data, the SBM is certainly among the most popular models for network data.

Within the SBM framework, the most important inferential task is that of recovering the community membership of the nodes from a *single* observation of the network. To solve this problem, in recent years researchers have proposed a variety procedures, which vary greatly in their degrees of statistical accuracy and computational complexity. See, in particular, modularity maximization [Newman and Girvan (2004)], likelihood methods [Amini et al. (2012), Bickel and Chen (2009), Celisse, Daudin and Pierre (2012), Choi, Wolfe and Airoldi (2012), Zhao, Levina and Zhu (2012)], method of moments [Anandkumar et al. (2013)], belief propagation [Decelle et al. (2011)], convex optimization [Chen, Sanghavi and Xu (2012)], spectral clustering [Balakrishnan et al. (2011), Fishkind et al. (2013), Jin (2012), Rohe, Chatterjee and Yu (2011), Sarkar and Bickel (2013)] and its variants [Chaudhuri, Chung and Tsiatas (2012), Coja-Oghlan (2010)] and spectral embeddings [Sussman et al. (2012), Lyzinski et al. (2013)].

Spectral clustering [see, e.g., von Luxburg (2007)] is arguably one of the most widely used methods for community recovery. Broadly speaking, this procedure first performs an eigen-decomposition of the adjacency matrix or the graph Laplacian. Then the community membership is inferred by applying a clustering algorithm, typically k-means, to the (possibly normalized) rows of the matrix formed by the first few leading eigenvectors. Spectral clustering is easier to implement and computationally less demanding than many other methods, most of which amount to computationally intractable combinatorial searches. From a theoretical standpoint, spectral clustering has been shown to enjoy good theoretical properties in denser stochastic block models where the average degree grows faster than $\log n$; see, for example, Jin (2012), Rohe, Chatterjee and Yu (2011), Sarkar and Bickel (2013). In addition, spectral clustering has been empirically observed to yield good performance even in sparser regimes. For example, it is recommended as the initial solution for a search based procedure in Amini et al. (2012). In computer science literature, spectral clustering is also a standard procedure for graph partitioning and for solving the planted partition model, a special case of the SBM [see, e.g., Ng et al. (2002)].

Despite its popularity and simplicity, the theoretical properties of spectral clustering are still not well understood in sparser SBM settings where the magnitude of the maximum expected node degree can be as small as $\log n$. This regime of sparsity is in fact not covered by existing analyses of the performance of spectral clustering for community recovery, which postulate a denser network. Indeed, Fishkind et al. (2013), Rohe, Chatterjee and Yu (2011) require the expected node degree to be almost linear in n, while Jin (2012) requires polynomial growth. Analogous conditions can be found elsewhere; see, for example, Sussman et al. (2012) and Balakrishnan et al. (2011).

In this paper, we derive new error bounds for spectral clustering for the purpose of community recovery in moderately sparse stochastic block models and degree corrected stochastic block models [see, e.g., Karrer and Newman (2011)], where the maximum expected node degree is of order $\log n$ or higher. Our main contribution is to show that the most basic form of spectral clustering is successful in recovering the latent community memberships under conditions on the network sparsity that are weaker than the ones used in most of literature. Our results yield some sharpening of existing analyses of spectral clustering for community recovery, and provide a theoretical justification for the effectiveness of this procedure in moderately sparse networks. We take note that there are competing methods yielding consistent community recovery under even milder conditions on the rate of growth of the node degrees, but they either rely on combinatorial methods that are computationally demanding [Bickel and Chen (2009)] or are guaranteed to be successful provided that they are given good starting points [Amini et al. (2012)], which are typically unknown. Other computationally efficient procedures with strong theoretical guarantees, which include in particular the ones proposed and analyzed in Channarond, Daudin and Robin (2012), Chen, Sanghavi and Xu (2012), McSherry (2001), Sarkar and Bickel (2013), require instead the degrees to be of larger order than $\log n$. More detailed comparisons with some of these contributions will be given after the statement of main results as more technical background is introduced. Finally, it is also known that in the ultra-sparse case, where the maximum degree is of order $O(1)$, consistent community recovery is impossible and one can only hope to recover the communities up to a constant fraction [see Coja-Oghlan (2010), Decelle et al. (2011), Krzakala et al. (2013), Massoulie (2013), Mossel, Neeman and Sly (2012, 2013)].

The contributions of this paper are as follows. We prove that a simplest form of spectral clustering, consisting of applying approximate k-means algorithms to the rows of the matrix formed by the leading eigenvectors of the adjacency matrix, allows to recover the community membeships of all but a vanishing fraction of the nodes in stochastic block models with expected degree as small as $\log n$, with high probability. We also extend this result to degree corrected stochastic block models by analyzing an approximate spherical k-median spectral clustering algorithm. The algorithms we consider are among the most practical and computationally affordable procedures available. Yet the theoretical guarantees we provide hold under rather general assumptions of sparsity that are weaker than the ones used in algorithms of similar complexity. Our arguments extend those in Rohe, Chatterjee and Yu (2011) and Jin (2012) by combining a principal subspace perturbation analysis (Lemma 5.1), a deterministic performance guarantee of approximate k-means clustering (Lemma 5.3) and a sharp bound on the spectrum of binary random matrices (Theorem 5.2), which may be of independent interest. These techniques give sharper results under weaker conditions. In particular, the subspace perturbation analysis allows us to avoid the individual eigengap condition. On the other hand, the spectral bound gives a better large deviation result that cannot be obtained by the matrix Bernstein inequality [Chung and Radcliffe (2011), Tropp (2012)] and leads to a simple extension to the degree corrected stochastic block model.

The article is organized as follows. In Section 2 we give formal introduction to the stochastic block model and spectral clustering. The main results are presented and compared to related works in Section 3 for regular SBM's and in Section 4 for degree corrected block models. Section 5 presents the proofs of main results, including a general, highly modular scheme of analyzing performance of spectral clustering algorithms. Concluding remarks are given in Section 6.

Notation. For a matrix M and index sets $\mathcal{I}, \mathcal{J} \subseteq [n]$, let $M_{\mathcal{I}*}$ and $M_{*\mathcal{J}}$ be the submatrix of M consisting the corresponding rows and columns. Let $\mathbb{M}_{n,K}$ be the collection of all $n \times K$ matrices where each row has exactly one 1 and $(K-1)$ 0's. For any $\Theta \in \mathbb{M}_{n,K}$, we call Θ a *membership matrix*, and the community membership of a node i is denoted by $g_i \in \{1, \dots, K\}$, which satisfies $\Theta_{ig_i} = 1$. Let $G_k = G_k(\Theta) = \{1 \le i \le n : g_i = k\}$ and $n_k = |G_k|$ for all $1 \le k \le K$. Let $n_{\min} = \min_{1 \le k \le K} n_k$, $n_{\max} = \max_{1 \le k \le K} n_k$, and $n'_{\max}$ be the second largest community size. We use $\|\cdot\|$ to denote both the Euclidean norm of a vector and the spectral norm of a matrix. $\|M\|_F = (\text{trace}(M^T M))^{1/2}$ denotes the Frobenius norm of a matrix M. The ℓ_0 norm $\|M\|_0$ simply counts the number of nonzero entries in M. For any square matrix M, $\text{diag}(M)$ denotes the matrix obtained by setting all off-diagonal entries of M to 0. For two sequences of real numbers $\{x_n\}$ and $\{y_n\}$, we will write $x_n = o(y_n)$ if $\lim_n x_n/y_n = 0$, $x_n = O(y_n)$ if $|x_n/y_n| \le C$ for all n and some positive C and $x_n = \Omega(y_n)$ if $|x_n/y_n| > C$ for all n and some positive C.

2. Preliminaries.

2.1. *Model setup.* A stochastic block model with n nodes and K communities is parameterized by a pair of matrices (Θ, B), where $\Theta \in \mathbb{M}_{n,K}$ is the membership matrix and $B \in \mathbb{R}^{K \times K}$ is a symmetric *connectivity matrix*. For each node i, let g_i $(1 \le g_i \le K)$ be its community label, such that the ith row of Θ is 1 in column g_i and 0 elsewhere. On the other hand, the entry $B_{k\ell}$ in B is the edge probability between any node in community k and any node in community ℓ. Given (Θ, B), the adjacency matrix $A = (a_{ij})_{1 \le i,j \le n}$ is generated as

$$a_{ij} = \begin{cases} \text{independent Bernoulli}(B_{g_i g_j}), & \text{if } i < j, \\ 10, & \text{if } i = j, \\ a_{ji}, & \text{if } i > j. \end{cases}$$

The goal of community recovery is to recover the membership matrix Θ up to column permutations. Throughout this article, we assume that the number of communities, K, is known. For an estimate $\widehat{\Theta} \in \mathbb{M}_{n,K}$ of the node memberships, we consider two measures of estimation error. The first one is an overall relative error

$$L(\widehat{\Theta}, \Theta) = n^{-1} \min_{J \in E_K} \|\widehat{\Theta} J - \Theta\|_0,$$

where E_K is the set of all $K \times K$ permutation matrices. Because both $\widehat{\Theta}J$ and Θ are membership matrices, we have $\|\widehat{\Theta}J - \Theta\|_0 = \|\widehat{\Theta}J - \Theta\|_F^2$. This quantity measures the overall proportion of mis-clustered nodes.

The other performance criterion measures the worst case relative error over all communities:

$$\tilde{L}(\widehat{\Theta}, \Theta) = \min_{J \in E_K} \max_{1 \le k \le K} n_k^{-1} \|(\widehat{\Theta}J)_{G_k*} - \Theta_{G_k*}\|_0.$$

It is obvious that $0 \le L(\widehat{\Theta}, \Theta) \le \tilde{L}(\widehat{\Theta}, \Theta) \le 2$. Thus, $\tilde{L}$ is a stronger criterion than L in that it requires the estimator to do well for all communities, while an estimator $\widehat{\Theta}$ with small $L(\widehat{\Theta}, \Theta)$ may have large relative errors for some small communities.

2.2. *Spectral clustering.* Spectral clustering is a simple method for community recovery [Jin (2012), Rohe, Chatterjee and Yu (2011), von Luxburg (2007)]. In a SBM, the heuristic of spectral clustering is to relate the eigenvectors of A to those of $P := \Theta B \Theta^T$ using the fact that $\mathbb{E}(A) = P - \operatorname{diag}(P)$. Let $P = UDU^T$ be the eigen-decomposition of P with $U^T U = I_K$ and $D \in \mathbb{R}^{K \times K}$ diagonal, then it is easy to see that U has only K distinct rows since P has only K distinct rows. Under mild conditions, it is also the case that two nodes are in the same community if and only if their corresponding rows in U are the same. This is formally stated in the following lemma.

LEMMA 2.1 (Basic eigen-structure of SBMs). *Let the pair* (Θ, B) *parametrize a SBM with* K *communities, where* B *is full rank. Let* UDU^T *be the eigen-decomposition of* $P = \Theta B \Theta^T$. *Then* $U = \Theta X$ *where* $X \in \mathbb{R}^{K \times K}$ *and* $\|X_{k*} - X_{\ell*}\| = \sqrt{n_k^{-1} + n_\ell^{-1}}$ *for all* $1 \le k < \ell \le K$.

PROOF. Let $\Delta = \operatorname{diag}(\sqrt{n_1}, \dots, \sqrt{n_K})$ then

$$P = \Theta B \Theta = \Theta \Delta^{-1} \Delta B \Delta (\Theta \Delta^{-1})^T. \tag{2.1}$$

It is straightforward to verify that $\Theta \Delta^{-1}$ is orthonormal. Let $ZDZ^T = \Delta B \Delta$ be the eigen-decomposition of $\Delta B \Delta$. Thus, we have $P = UDU^T$ where $U = \Theta \Delta^{-1} Z$. The claim follows by letting $X = \Delta^{-1} Z$ and realizing that the rows of $\Delta^{-1} Z$ are perpendicular to each other and the kth row has length $\|(\Delta Z)_{k*}\| = \sqrt{1/n_k}$. □

Based on this observation, spectral clustering tries to estimate U and its row clustering using a spectral decomposition of A. The intuition for the procedure is as follows. Consider the difference between A and P:

$$A - P = (A - \mathbb{E}(A)) - \operatorname{diag}(P),$$

which is a symmetric noise matrix plus a diagonal matrix. Intuitively, the eigenvectors of A will be close to those of P because the eigenvalues of P scales linearly

```
Algorithm 1: Spectral clustering with approximate $k$-means
Input: Adjacency matrix $A$; number of communities $K$; approximation parameter $\varepsilon$.
Output: Membership matrix $\widehat{\Theta} \in \mathbb{M}_{n,K}$.
1. Calculate $\widehat{U} \in \mathbb{R}^{n\times K}$ consisting of the leading $k$ eigenvectors (ordered
   in absolute eigenvalue) of $A$.
2. Let $(\widehat{\Theta}, \widehat{X})$ be an $(1+\varepsilon)$-approximate solution to the $k$-means problem
   (2.3) with $K$ clusters and input matrix $\widehat{U}$.
3. Output $\widehat{\Theta}$.
```

with n while the noise matrix $(A - \mathbb{E}(A))$ has operator norm on the scale of $\sqrt{n}$ and $\operatorname{diag}(P)$ is like a constant. Therefore, letting $A = \widehat{U}\widehat{D}\widehat{U}^T$ be the K-dimensional eigen-decomposition of A corresponding to the K largest absolute eigenvalues, we can see that $\widehat{U}$ should have roughly K distinct rows because they are slightly perturbed versions of the rows in U. Then one should be able to obtain a good community partition by applying a clustering algorithm on the rows of $\widehat{U}$. In this paper we consider the k-means clustering, defined as

$$(2.2) \qquad (\widehat{\Theta}, \widehat{X}) = \arg \min_{\Theta\in\mathbb{M}_{n,K}, X\in\mathbb{R}^{K\times K}} \|\Theta X - \widehat{U}\|_F^2.$$

It is known that finding a global minimizer for the k-means problem (2.2) is NP-hard [see, e.g., Aloise et al. (2009)]. However, efficient algorithms exist for finding an approximate solution whose value is within a constant fraction of the optimal value [Kumar, Sabharwal and Sen (2004)]. That is, there are polynomial time algorithms that find

$$(2.3) \qquad \begin{aligned} &(\widehat{\Theta}, \widehat{X}) \in \mathbb{M}_{n,K} \times \mathbb{R}^{K\times K} \\ \text{s.t.} \quad &\|\widehat{\Theta}\widehat{X} - \widehat{U}\|_F^2 \le (1+\varepsilon) \min_{\Theta\in\mathbb{M}_{n,K}, X\in\mathbb{R}_{K\times K}} \|\Theta X - \widehat{U}\|_F^2. \end{aligned}$$

The spectral clustering algorithm we consider here is summarized in Algorithm 1.

2.3. *Sparsity scaling.* Real-world large scale networks are usually sparse, in the sense that the number of edges from a node (the node degree) are very small compared to the total number of nodes. Generally speaking, community recovery is hard when data is sparse. As a result, an important criterion of evaluating a community recovery method is its performance under different levels of sparsity (usually measured in the error rate as a function of the average/maximum degree). The following prototypical example exemplifies well the roles played by network sparsity as well as other model parameters in determining the hardness of community recovery.

EXAMPLE 2.2. Consider a SBM with K communities parameterized by (Θ, B) where

$$B = \alpha_n B_0; \qquad B_0 = \lambda I_K + (1-\lambda)\mathbf{1}_K \mathbf{1}_K^T, \qquad 0 < \lambda < 1, \tag{2.4}$$

I_K is the $K \times K$ identity matrix, and $\mathbf{1}_K$ is the $K \times 1$ vector of 1's.

Example 2.2 assumes that the edge probability between any pair of nodes depends only on whether they belong to the same community. In particular, the edge probability is α_n within community and $\alpha_n(1-\lambda)$ between community. The quantity λ reflects the relative difference in connectivity between communities and within communities. The network sparsity is captured by α_n, where $n\alpha_n$ provides an upper bound on the average (and maximum in this example) expected node degree. It can be easily seen that if α_n or λ are close to 0 then it is hard to identify communities.

The hardness of community reconstruction also depends on the number of communities and the community size imbalance. For example, the famous planted clique problem concerns community recovery under a SBM with $K = 2$ and

$$B = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}. \tag{2.5}$$

In the planted clique problem, it is known that community recovery is easy if $n_1 \ge c\sqrt{n}$ for a constant c [see Deshpande and Montanari (2013) and references therein] and on the other hand no polynomial time algorithms have been found to succeed when $n_1 = o(\sqrt{n})$.

REMARK. The primary concern of this paper is the effect of α_n on the performance of spectral clustering. Nevertheless, our results explicitly keep track of other quantities such as K, λ, $n_{\max}$ and $n_{\min}$, all of which are allowed to change with n in a nontrivial manner. The dependence of recovery error bound on some of these quantities, such as K and λ, is concerned by some authors, such as Anandkumar et al. (2013), Chaudhuri, Chung and Tsiatas (2012), Chen, Sanghavi and Xu (2012). For ease of readability, we do not always make this dependence on n explicit in our notation.

3. Stochastic block models.

Our main result provides an upper bound on relative community reconstruction error of spectral clustering for a SBM (Θ, B) in terms of several model parameters.

THEOREM 3.1. *Let A be an adjacency matrix generated from a stochastic block model (Θ, B). Assume that $P = \Theta B \Theta^T$ is of rank K, with smallest absolute nonzero eigenvalue at least γ_n and* $\max_{k,\ell} B_{k\ell} \le \alpha_n$ *for some $\alpha_n \ge \log n/n$. Let*

$\widehat{\Theta}$ be the output of spectral clustering using $(1+\varepsilon)$-approximate k-means (Algorithm 1). There exists an absolute constant $c>0$, such that, if

$$(2+\varepsilon)\frac{Kn\alpha_n}{\gamma_n^2}<c, \tag{3.1}$$

then, with probability at least $1-n^{-1}$, there exist subsets $S_k\subset G_k$ for $k=1,\dots,K$, and a $K\times K$ permutation matrix J such that $\widehat{\Theta}_{G*}J=\Theta_{G*}$, where $G=\bigcup_{k=1}^K(G_k\setminus S_k)$, and

$$\sum_{k=1}^{n}\frac{|S_k|}{n_k}\le c^{-1}(2+\varepsilon)\frac{Kn\alpha_n}{\gamma_n^2}. \tag{3.2}$$

The proof of Theorem 3.1, given in Section 5, is modular, and can be derived from several relatively independent lemmas.

The sets S_k $(1\le k\le K)$ consist of nodes in G_k for which the clustering correctness cannot be guaranteed. The permutation matrix J in the above theorem leads to an upper bound on reconstruction error $\tilde{L}(\widehat{\Theta},\Theta)$ [and hence on $L(\widehat{\Theta},\Theta)$] through equation (3.2).

Condition (3.1) specifies the range of model parameters (K,n,γ_n,α_n) for which the result is applicable. It is included only for technical reasons, because it holds whenever the bound in (3.2) vanishes and, therefore, implies consistency. In particular, as discussed after Corollary 3.2, we have $Kn\alpha_n/\gamma_n^2=o(1)$ in many interesting cases. The constant c in (3.1) can be written as $c=1/(64C^2)$ where C is an absolute constant defined in Theorem 5.2 and can be explicitly tracked in the proof presented in the supplementary material [Lei and Rinaldo (2014)]. The assumption of $\alpha_n\ge\log n/n$ can be changed to $\alpha_n\ge c_0\log n/n$ for any $c_0>0$, and also the probability bound $1-n^{-1}$ can be changed to $1-n^{-r}$ for any $r>0$, with a different constant $c=c(c_0,r)$ in (3.1) and (3.2).

While Theorem 3.1 provides a general error bound for spectral clustering, the quantities involved are not in the most transparent form. For example, the bound does not clearly reflect the intuition that the error should increase when α_n decreases. This is because the quantity γ_n contains the parameter α_n. Also the dependence on the community size imbalance as well as the community separation (which corresponds to the parameter λ in Example 2.2) remains unclear. The next corollary illustrates the error bound in terms of these model parameters.

COROLLARY 3.2. *Let A be an adjacency matrix from the SBM (Θ,B), where $B=\alpha_nB_0$ for some $\alpha_n\ge\log n/n$ and with B_0 having minimum absolute eigenvalue $\ge\lambda>0$ and $\max_{k\ell}B_0(k,\ell)=1$. Let $\widehat{\Theta}$ be the output of spectral clustering using $(1+\varepsilon)$-approximate k-means (Algorithm 1). Then there exists an absolute constant c such that if*

$$(2+\varepsilon)\frac{Kn}{n_{\min}^2\lambda^2\alpha_n}<c \tag{3.3}$$

then with probability at least $1 - n^{-1}$,

$$\tilde{L}(\widehat{\Theta}, \Theta) \le c^{-1}(2+\varepsilon)\frac{Kn}{n_{\min}^2\lambda^2\alpha_n}$$

and

$$L(\widehat{\Theta}, \Theta) \le c^{-1}(2+\varepsilon)\frac{Kn'_{\max}}{n_{\min}^2\lambda^2\alpha_n}.$$

In the special case of a balanced community sizes [i.e., $n_{\max}/n_{\min} = O(1)$] and constant λ, if $\alpha_n = \Omega(\log n/n)$, then $L(\widehat{\Theta}, \Theta) = O_P(K^2(n\alpha_n)^{-1}) = O_P(K^2/\log n)$. Thus $L(\widehat{\Theta}, \Theta) = o_P(1)$ if $K = o(\sqrt{\log n})$. This improves the results in Rohe, Chatterjee and Yu (2011) where α_n needs to be of order $1/\log n$ for a similar result.

In Example 2.2, the smallest nonzero eigenvalue of B_0 is λ. Recall that λ is the relative difference of within- and between-community edge probabilities. Corollary 3.2 then implies that when this relative difference stays bounded away from zero, the communities can be consistently recovered by simple spectral clustering as long as the expected node degrees are no less than $\log n$. On the other hand, when α_n is constant and $\lambda = \lambda_n$ varies with n, spectral clustering can recover the communities when the relative edge probability gap grows faster than $1/\sqrt{n}$.

In the planted clique problem, $L(\widehat{\Theta}, \Theta)$ has limited meaning because a trivial clustering putting all nodes in one cluster achieves $L(\widehat{\Theta}, \Theta) = 2n_{\min}/n$ which is $o(1)$ in the most interesting regime. Therefore, it makes more sense to consider $\tilde{L}(\widehat{\Theta}, \Theta)$. Now $B_0 = B$ is given by (2.5), with minimum eigenvalue > 0.19. Applying Corollary 3.2 with $K = 2$, $\lambda = 0.19$, $\alpha_n = 1$, and any fixed $\varepsilon > 0$, we have

$$\tilde{L}(\widehat{\Theta}, \Theta) < c'\frac{n}{n_{\min}^2},$$

provided that $c'n/n_{\min}^2 < 1$, where c' is a different absolute constant. Therefore, when $n_{\min} \ge \sqrt{an}$ for some $a > c'$, $\widehat{\Theta}$ recovers the hidden clique with a relative error no larger than c'/a. Thus, our result reaches the well believed computation barrier [up to constant factor, see Deshpande and Montanari (2013) and references therein] of the planted clique problem.

There are spectral methods other than spectral clustering that can provide consistent community recovery. One such well-known method is the procedure analyzed by McSherry (2001). The planted partition problem in that setting corresponds to the problem of recovering the community memberships in the SBM. To simplify the presentation and focus on the dependence of network sparsity, we consider the SBM in Example 2.2 with two equal-sized communities and a constant $\lambda \in (0, 1)$. According to Theorem 4 in McSherry (2001), that method can recover the true communities with probability at least $1 - n^{-1}$ provided that, after some simplification,

$$\lambda^2\alpha_n^2 n > c\sigma_n^2 \log n \quad \text{and} \quad \sigma_n^2 > (\log n)^6/n, \tag{3.4}$$

for some constant c, where σ_n^2 is an upper bound on the maximal variance of the edges. Therefore, the condition (3.4) implies that $\alpha_n > \sqrt{c}\lambda^{-1}(\log n)^{3.5}/n$, which is stronger than the condition in our Corollary 3.2.

4. Degree corrected stochastic block models. The degree corrected block model [DCBM, Karrer and Newman (2011)] extends the standard SBM by introducing node specific parameters to allow for varying degrees even within the same community. A DCBM is parameterized by a triplet (Θ, B, ψ), where, in addition to the membership matrix Θ and connectivity matrix B, the vector $\psi \in \mathbb{R}^n$ is included to model additional variability of the edge probabilities at the node level. Given (Θ, B, ψ), the edge probability between nodes i and j is $\psi_i \psi_j B_{g_i g_j}$ (recall that g_i is the community label of node i). Similar to the SBM, the DCBM also assumes independent edge formation given (Θ, B, ψ). The inclusion of ψ raises an issue of identifiability. So we assume that $\max_{i \in G_k} \psi_i = 1$ for all $k = 1, \dots, K$. The SBM can be viewed as a special case of DCBM with $\psi_i = 1$ for all i. The DCBM greatly enhances the flexibility of modeling degree heterogeneity and is able to fit network data with arbitrary degree distribution. Successful application and theoretical developments can be found in Zhao, Levina and Zhu (2012) for likelihood methods, and in Chaudhuri, Chung and Tsiatas (2012), Jin (2012) for spectral methods.

Additional notation about the degree heterogeneity. Let ϕ_k be the $n \times 1$ vector that agrees with ψ on G_k and zero otherwise. Define $\tilde{\phi}_k = \phi_k / \|\phi_k\|$ and $\tilde{\psi} = \sum_{k=1}^{K} \tilde{\phi}_k$. Let $\tilde{\Theta}$ be a normalized membership matrix such that $\tilde{\Theta}(i,k) = \tilde{\psi}_i$ if $i \in G_k$ and $\tilde{\Theta}(i,k) = 0$ otherwise. We also define *effective community size* $\tilde{n}_k := \|\phi_k\|^2$. Let $\tilde{n}_{\min} = \min_k \tilde{n}_k$ and $\tilde{n}_{\max} = \max_k \tilde{n}_k$.

The spectral clustering heuristic can be extended to DCBMs by considering the eigen-decomposition $P = UDU^T$ where $P = \operatorname{diag}(\psi)\Theta B \Theta^T \operatorname{diag}(\psi)$. Now the matrix U may have more than K distinct rows due to the effect of ψ. However, the rows of U point to at most K distinct directions [Jin (2012)]. The following lemma is the analogue of Lemma 2.1 for DCBMs.

LEMMA 4.1 (Spectral structure of mean matrix in DCBM). *Let UDU^T be the eigen-decomposition of $P = \operatorname{diag}(\psi)\Theta B \Theta^T \operatorname{diag}(\psi)$ in a DCBM parameterized by (Θ, B, ψ). Then there exists a $K \times K$ orthogonal matrix H such that*

$$U_{i*} = \tilde{\psi}_i H_{k*} \qquad \forall 1 \le k \le K, i \in G_k.$$

PROOF. First, realize that $\operatorname{diag}(\psi)\Theta = \tilde{\Theta}\Psi$, where $\Psi = \operatorname{diag}(\|\phi_1\|, \dots, \|\phi_K\|)$.

$$P = \operatorname{diag}(\psi)\Theta B \Theta^T \operatorname{diag}(\psi) = \tilde{\Theta}\Psi B \Psi \tilde{\Theta}^T = \tilde{\Theta} H D (\tilde{\Theta} H)^T, \tag{4.1}$$

where $\Psi B \Psi = HDH^T$ is the eigen-decomposition of $\Psi B \Psi$. Note that $\tilde{\Theta}^T \tilde{\Theta} = I_K$ so $\tilde{\Theta} H D (\tilde{\Theta} H)^T$ is an eigen-decomposition of P. □

As a result, finding the true community partition corresponds to clustering the *directions* of the row vectors in U, where some form of normalization must be employed in order to filter out the nuisance parameter ψ. In particular, we consider *spherical clustering*, which looks for a cluster structure among the rows of a normalized matrix U' with $U'_{i*} = U_{i*}/\|U_{i*}\|$.

In addition to the overall sparsity, the difficulty of community recovery in a DCBM is also affected by small entries of ψ. Intuitively, if $\psi_i \approx 0$, then it is hard to identify the community membership of node i because few edges are observed for this node. However, the interaction between small entries of ψ and the overall network sparsity (the maximum/average degree) has not been well understood. In the analysis of profile likelihood methods, Zhao, Levina and Zhu (2012) assume that the entries of ψ are fixed constants. In spectral clustering, Jin (2012) allows milder conditions on ψ but needs the average degree to be polynomial in n.

Our analysis uses the following quantity as a summarizing measure of node heterogeneity in each community G_k:

$$\nu_k := n_k^{-2} \sum_{i \in G_k} \tilde{\psi}_i^{-2}, \qquad k = 1, 2, \dots, K.$$

By definition $\nu_k \in [1, \infty)$ and a larger ν_k indicates a stronger heterogeneity in the kth community. On the other hand, $\nu_k = 1$ indicates within-community homogeneity ($\psi_i = 1$ for all $i \in G_k$).

The argument developed for SBMs in previous sections can be extended to cover very general degree corrected models. In particular, let $\widehat{U} \in \mathbb{R}^{n \times K}$ consist the K leading eigenvectors of A. We consider the following spherical k-median spectral clustering:

$$\text{minimize}_{\Theta \in \mathbb{M}_{n,K}, X \in \mathbb{R}^{K \times K}} \|\Theta X - \widehat{U}'\|_{2,1}, \tag{4.2}$$

where $\widehat{U}'$ is the row-normalized version of $\widehat{U}$ and $\|M\|_{2,1} = \sum_{i=1} \|M_{i*}\|$ is the matrix (2, 1)-norm. We will not require to solve (4.2) exactly but instead we consider a $(1 + \varepsilon)$ approximation $(\widehat{\Theta}, \widehat{X})$ to the k-median problem, which can be solved in polynomial time when $\varepsilon > \sqrt{3}$ [Charikar et al. (1999), Li and Svensson (2013)]. The practical procedure will also take care of the possible zero rows in $\widehat{U}$ and is described in detail in Algorithm 2.

4.1. *Analysis of spherical k-median spectral clustering for DCBM.*

We have the following main theorem for spherical k-median spectral clustering in DCBMs. It is proved in Appendix A.3.

THEOREM 4.2 (Main result for DCBM). *Consider a DCBM (Θ, B, ψ) with K communities, where $P = \operatorname{diag}(\psi)\Theta B \Theta^T \operatorname{diag}(\psi)$ has rank K, the smallest nonzero absolute eigenvalue at least γ_n, and the maximum entry bounded from*

```
Algorithm 2: Spherical $k$-median spectral clustering
Input: Adjacency matrix $A$; number of communities $K$; approximation parameter $\varepsilon$.
Output: Membership matrix $\widehat{\Theta} \in \mathbb{M}_{n,K}$.
1. Calculate $\widehat{U} \in \mathbb{R}^{n\times K}$ consisting of the leading $k$ eigenvectors (ordered in absolute eigenvalue) of $A$.
2. Let $I_+ = \{i : \|\widehat{U}_{i*}\| > 0\}$ and $\widehat{U}^+ = (\widehat{U}_{I_+*})$.
3. Let $\widehat{U}'$ be row-normalized version of $\widehat{U}^+$.
4. Let $(\widehat{\Theta}^+, \widehat{X})$ be an $(1+\varepsilon)$-approximate solution to the $k$-median problem with $K$ clusters and input matrix $\widehat{U}'$.
5. Output $\widehat{\Theta}$ with $\widehat{\Theta}_{i*}$ being the corresponding row in $\widehat{\Theta}^+$ if $i \in I_+$, and $\widehat{\Theta}_{i*} = (1, 0, \dots, 0)$ if $i \notin I_+$.
```

above by $\alpha_n \geq \log n/n$. There exists an absolute constant $c > 0$ such that if

$$(2.5+\varepsilon)\frac{\sqrt{Kn\alpha_n}}{\gamma_n} < c\frac{n_{\min}}{\sqrt{\sum_{k=1}^K n_k^2 \nu_k}} \tag{4.3}$$

then, with probability at least $1 - n^{-1}$,

$$L(\widehat{\Theta}, \Theta) \leq c^{-1}(2.5+\varepsilon)\sqrt{\sum_{k=1}^K n_k^2 \nu_k}\frac{\sqrt{K\alpha_n}}{\gamma_n\sqrt{n}}. \tag{4.4}$$

REMARK. The constant c equals $1/(8C)$ where C is the universal constant in Theorem 5.2. The condition on α_n and probability guarantee can also be changed to $\alpha_0 \geq c_0 \log n/n$ and $1 - n^{-r}$, respectively, with a different constant $c = c(c_0, r)$ in equations (4.3) and (4.4).

Theorem 4.2 immediately implies a counterpart of Corollary 3.2 under more explicit scaling of the model parameters.

COROLLARY 4.3. *Let A be an adjacency matrix from DCBM (Θ, B, ψ), such that $B = \alpha_n B_0$ for some $\alpha_n \geq \log n/n$ where B_0 has minimum absolute eigenvalue $\lambda > 0$ and $\max_{k\ell} B_0(k,\ell) = 1$. Let $(\widehat{\Theta}, \widehat{X})$ be an $(1+\varepsilon)$-approximate solution to the spherical k-median algorithm (Algorithm 2). There exists an absolute constant c such that if*

$$(2.5+\varepsilon)\frac{\sqrt{Kn}}{\tilde{n}_{\min}\lambda\sqrt{\alpha_n}} < c\frac{n_{\min}}{\sqrt{\sum_{k=1}^K n_k^2 \nu_k}},$$

then, with probability at least $1 - n^{-1}$,

$$L(\widehat{\Theta}, \Theta) \le c^{-1}(2.5 + \varepsilon)\frac{\sqrt{K}}{\tilde{n}_{\min}\lambda\sqrt{n\alpha_n}}\sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}.$$

Comparing with Theorem 3.1 and Corollary 3.2, the results for DCBM are different in two major aspects. First, the DCBM condition (4.3) involves the term $n_{\min}^2/\sum_{k=1}^K n_k^2\nu_k$ which is smaller than 1 (indeed smaller than $1/K$). This makes (4.3) more stringent than (3.1). Also the upper bound on $L(\widehat{\Theta}, \Theta)$ is different in the same manner. Furthermore, the argument used to prove Theorem 4.2 is not likely to provide a sharp upper bound on $\tilde{L}(\widehat{\Theta}, \Theta)$. We believe this has to do with the additional normalization step used in the spherical k-median algorithm as well as the specific strategy used in our proof.

To better understand this result, consider Example 2.2 with balanced community size: $n_{\max}/n_{\min} = O(1)$. To work with a DCBM, assume in addition that the node degree vector ψ has comparable degree heterogeneity across communities: $c_1\nu \le \nu_k \le c_2\nu$ for constants c_1, c_2. Then Corollary 4.3 implies an overall relative error rate

$$L(\widehat{\Theta}, \Theta) = O_P\left(\frac{\sqrt{\nu}}{\tilde{n}_{\min}\lambda\sqrt{n\alpha_n}}\right). \tag{4.5}$$

Several observations are worth mentioning. First, the error rate depends on ν, the degree heterogeneity measure, in a simple manner. Second, the community size $n_{\min}$ that appears in Corollary 3.2 is replaced by $\tilde{n}_{\min} = \min_k \|\phi_k\|$, the minimum effective sample size. Roughly speaking, $\tilde{n}_{\min} \asymp n_{\min}$ as long as a constant fraction of nodes have their ψ_i's bounded away from zero (but the rest should not be too small in order to keep ν small). Third, if there is no degree heterogeneity ($\nu_k \equiv 1$ and $\tilde{n}_{\min} = n_{\min}$), then the rate in (4.5) is the square root of that given by Corollary 3.2. This is due to the additional normalization step (which is not necessary since $\nu = 1$) involved in spherical k-median and the different argument used to analyze the spherical k-median algorithm. Moreover, the relative error can still be $o_P(1)$ even when α_n is as small as $\log n/n$, provided that $1/\nu$, $\tilde{n}_{\min}/n$, and λ stay bounded away from zero or approach zero sufficiently slowly.

Comparisons with existing work. There are relatively fewer results for community recovery in degree corrected block models that allow the maximum node degree to be of order $o(n)$. Chaudhuri, Chung and Tsiatas (2012) extended the method of McSherry (2001) to degree corrected block models. In the setting of Example 2.2 with equal community size, their main result (Theorems 2 and 3 in their paper) requires α_n to be at least of order $1/\sqrt{n}$. A similar requirement of a polynomial growth of expected average degree is implicitly imposed in Jin (2012), who first studied the performance of normalized k-means spectral clustering in degree corrected block models.

5. Proof of the main results. In this section, we present a general scheme to prove error bounds for spectral clustering. It contains the SBM as a special case and can be easily extended to the degree corrected block model. Our argument consists of three parts: (1) control the perturbation of principal subspaces for general symmetric matrices, (2) bound the spectrum of random binary matrices, and (3) error bound of k-mean and spherical k-median clustering.

5.1. *Principal subspace perturbation.* The first ingredient of our proof is to bound the difference between the eigenvectors of A and those of P, where A can be viewed as a noisy version of P.

LEMMA 5.1 (Principal subspace perturbation). *Assume that $P \in \mathbb{R}^{n\times n}$ is a rank K symmetric matrix with smallest nonzero singular value γ_n. Let A be any symmetric matrix and $\widehat{U}, U \in \mathbb{R}^{n\times K}$ be the K leading eigenvectors of A and P, respectively. Then there exists a $K \times K$ orthogonal matrix Q such that*

$$\|\widehat{U} - UQ\|_F \le \frac{2\sqrt{2K}}{\gamma_n}\|A - P\|.$$

Lemma 5.1 is proved in Appendix A.1, which is based on an application of the Davis–Kahan $\sin\Theta$ theorem [Theorem VII.3.1 of Bhatia (1997)]. The presence of a $K \times K$ orthonormal matrix Q in the statement of Lemma 5.1 is to take care of the situation where some leading eigenvalues have multiplicities larger than one. In this case, the eigenvectors are determined only up to a rotation.

5.2. *Spectral bound of binary symmetric random matrices.* The next theorem provides a sharp probabilistic upper bound on $\|A - P\|$ when A is a random adjacency matrix with $\mathbb{E}(a_{ij}) = p_{ij}$.

THEOREM 5.2 (Spectral bound of binary symmetric random matrices). *Let A be the adjacency matrix of a random graph on n nodes in which edges occur independently. Set $\mathbb{E}[A] = P = (p_{ij})_{i,j=1,\dots,n}$ and assume that $n \max_{ij} p_{ij} \le d$ for $d \ge c_0 \log n$ and $c_0 > 0$. Then, for any $r > 0$ there exists a constant $C = C(r, c_0)$ such that*

$$\|A - P\| \le C\sqrt{d}$$

with probability at least $1 - n^{-r}$.

This result does not follow conventional matrix concentration inequalities such as the matrix Bernstein inequality (which will only give $\sqrt{d\log n}$). Lu and Peng (2012) use a path counting technique in random matrix theory to prove a bound of the same order but require a maximal degree $d \ge c_0(\log n)^4$.

The proof of Theorem 5.2 is technically involved, as it uses combinatorial arguments in order to derive spectral bounds for sparse random matrices. Our proof is based on techniques developed by Feige and Ofek (2005) for bounding the second largest eigenvalue of an Erdös–Réyni random graph with edge probability d/n. The full proof is provided in Lei and Rinaldo (2014). Here we give a brief outline of the three major steps.

Step 1: *Discretization.* We first reduce controlling $\|A - P\|$ to the problem of bounding the supremum of $|x^T(A - P)y|$ over all pairs of vectors x, y in a finite set of grid points. For any given pair (x, y) in the grid, the quantity $x^T(A - P)y$ is decomposed into the sum of two parts. The first part corresponds to the small entries of both x and y, called *light pairs*, the other part corresponds to the larger entries of x or y, the *heavy pairs*.

Step 2: *Bounding the light pairs.* The next step is to use Bernstein's inequality and the union bound to control the contribution of the light pairs, uniformly over the points in the grid.

Step 3: *Bounding the heavy pairs.* In the final step, the contribution from the heavy pairs, which cannot be simply bounded by conventional Bernstein's inequality, will be bounded using a combinatorial argument on the event that the edge numbers in a collection of subgraphs do not deviate much from their expectation. A sharp large deviation bound for sums of independent Bernoulli random variables [Corollary A.1.10 of Alon and Spencer (2004)] is used to achieve better rate than standard Bernstein's inequality.

5.3. *Error bound of k-means/k-median on perturbed eigenvectors.* Spectral clustering (or spherical spectral clustering) applies a clustering algorithm to a matrix consisting of the eigenvectors of A, which is close (in view of Lemma 5.1 and Theorem 5.2) to a matrix whose rows can be perfectly clustered. We would like to bound the clustering error in terms of the closeness between the real input matrix $\widehat{U}$ and the ideal input matrix U.

The next lemma generalizes an argument used in Jin (2012) and provides an error bound for any $(1 + \varepsilon)$-approximate k-means solution.

LEMMA 5.3 (Approximate k-means error bound). *For $\varepsilon > 0$ and any two matrices $\widehat{U}, U \in \mathbb{R}^{n\times K}$ such that $U = \Theta X$ with $\Theta \in \mathbb{M}_{n,K}$, $X \in \mathbb{R}^{K\times K}$, let $(\widehat{\Theta}, \widehat{X})$ be a $(1 + \varepsilon)$-approximate solution to the k-means problem in equation* (2.2) *and $\bar{U} = \widehat{\Theta}\widehat{X}$. For any $\delta_k \le \min_{\ell\neq k} \|X_{\ell *} - X_{k*}\|$, define $S_k = \{i \in G_k(\Theta) : \|\bar{U}_{i*} - U_{i*}\| \ge \delta_k/2\}$ then*

$$\sum_{k=1}^{K} |S_k|\delta_k^2 \le 4(4 + 2\varepsilon)\|\widehat{U} - U\|_F^2. \tag{5.1}$$

Moreover, if

$$(16 + 8\varepsilon)\|\widehat{U} - U\|_F^2/\delta_k^2 < n_k \qquad \textit{for all } k, \tag{5.2}$$

then there exists a $K \times K$ *permutation matrix* J *such that* $\widehat{\Theta}_{G*} = \Theta_{G*} J$, *where* $G = \bigcup_{k=1}^{K} (G_k \setminus S_k)$.

Lemma 5.3 provides a performance guarantee for approximate k-means clustering under a deterministic Frobenius norm condition on the input matrix. As suggested by a referee, the proof of Lemma 5.3 shares some similarities with the proof of Theorem 3.1 in Awasthi and Sheffet (2012) [see also Kumar and Kannan (2010)], though our assumptions are slightly different. For completeness we provide a short and self-contained proof of Lemma 5.3 in Appendix A.2, giving explicit constant factors in the result.

5.4. *Proof of main results for SBM.* We first prove Theorem 3.1.

PROOF OF THEOREM 3.1. Combining Lemma 5.1 and Theorem 5.2, we obtain that, for some K-dimensional orthogonal matrix Q,

$$\|\widehat{U} - UQ\|_F \le \frac{2\sqrt{2K}}{\gamma_n}\|A - P\| \le \frac{2\sqrt{2K}}{\gamma_n} C\sqrt{n\alpha_n}, \tag{5.3}$$

with probability at least $1 - n^{-1}$, where C is the absolute constant involved in Theorem 5.2. (Notice that the term d in Theorem 5.2 becomes $n\alpha_n$ in the current setting.)

The main strategy for the rest of the proof is to apply Lemma 5.3 to $\widehat{U}$ and UQ. To that end, Lemma 2.1 implies that $UQ = \Theta X Q = \Theta X'$ where $\|X'_{k*} - X'_{\ell*}\| = \sqrt{\frac{1}{n_k} + \frac{1}{n_\ell}}$. As a result, we can choose $\delta_k = \sqrt{1/n_k + \frac{1}{\max\{n_\ell : \ell \neq k\}}}$ in Lemma 5.3 and hence $n_k \delta_k^2 \ge 1$ for all k. Using (5.3), a sufficient condition for (5.2) to hold is

$$(16 + 8\varepsilon) 8C^2 K \frac{n\alpha_n}{\gamma_n^2} \le 1 \le \min_{1 \le k \le K} n_k \delta_k^2, \tag{5.4}$$

so that (3.1) indeed implies (5.2) by setting $c = \frac{1}{64C^2}$. In detail, the choice of $\delta_k = 1/\sqrt{n_k}$ together with (5.1) yields that

$$\sum_{k=1}^{K} |S_k| \left(\frac{1}{n_k} + \frac{1}{\max\{n_\ell : \ell \neq k\}} \right) = \sum_{k=1}^{K} |S_k| \delta_k^2 \le 4(4 + 2\varepsilon) \|\widehat{U} - UQ\|_F^2,$$

which, combined with (5.3), gives (3.2):

$$\sum_{k=1}^{K} \frac{|S_k|}{n_k} \le 4(4 + 2\varepsilon) 8C^2 \frac{Kn\alpha_n}{\gamma_n^2} = c^{-1}(2 + \varepsilon) \frac{Kn\alpha_n}{\gamma_n^2}.$$

Since Lemma 5.3 ensures that the membership is correctly recovered outside of $\bigcup_{1 \le k \le K} S_k$, the claim follows. □

PROOF OF COROLLARY 3.2. It is easy to see, for example, from (2.1), that in this specific stochastic block model setting, $\gamma_n = n_{\min}\alpha_n\lambda$. Then the proof of Theorem 3.1 applies with $\gamma_n = n_{\min}\alpha_n\lambda$ and gives

$$\sum_{k=1}^{K} |S_k|\left(\frac{1}{n_k} + \frac{1}{\max\{n_\ell : \ell \neq k\}}\right) \leq 64C^2(2+\varepsilon)\frac{Kn}{n_{\min}^2\lambda^2\alpha_n},$$

which implies that

$$\tilde{L}(\widehat{\Theta}, \Theta) \leq \max_{1\leq k\leq K} \frac{|S_k|}{n_k} \leq \sum_{1\leq k\leq K} \frac{|S_k|}{n_k} \leq 64C^2(2+\varepsilon)\frac{Kn}{n_{\min}^2\lambda^2\alpha_n},$$

and, recalling that $n'_{\max}$ is the second largest community size,

$$L(\widehat{\Theta}, \Theta) \leq \frac{1}{n}\sum_{k=1}^{K} |S_k| \leq 64C^2(2+\varepsilon)\frac{Kn'_{\max}}{n_{\min}^2\lambda^2\alpha_n}. \qquad \square$$

6. Concluding remarks.

The analysis in this paper applies directly to the eigenvectors of the adjacency matrix, by combining tools in subspace perturbation and spectral bounds of binary random graphs. In the literature, spectral clustering using the graph Laplacian or its variants is very popular and can sometimes lead to better empirical performance [Rohe, Chatterjee and Yu (2011), Sarkar and Bickel (2013), von Luxburg (2007)]. An important future work would be to extend some of the results and techniques in this paper to spectral clustering using the graph Laplacian. The graph Laplacian normalizes the adjacency matrix by the node degree, which can introduce extra noise if the network is sparse and many node degrees are small. In several recent works, Chaudhuri, Chung and Tsiatas (2012), Qin and Rohe (2013) studied graph Laplacian based spectral clustering with regularization, where a small constant is added to all node degrees prior to the normalization. Further understanding the bias-variance trade off would be both important and interesting.

For degree corrected block models, regularization methods may also lead to error bounds with better dependence on small entries of ψ. The intuition is that ν_k can be very large even when only one ψ_i is very close to zero. In this case, one should be able to simply discard nodes like this and work on those with large enough degrees. Finding the correct regularization to diminish the effect of small-degree nodes and analyzing the new algorithm will be pursued in future work.

This paper aims at understanding the performance of spectral clustering in stochastic block models. While our main focus is the performance of spectral clustering as the network sparsity changes, the resulting error bounds explicitly keep track of five independent model parameters $(K, \alpha_n, \lambda, n_{\min}, n_{\max})$. Existing results usually develop error bounds depending on a subset of these parameters, keeping others as constant [see, e.g., Bickel and Chen (2009), Chen, Sanghavi and Xu (2012), Zhao, Levina and Zhu (2012)]. In the planted clique model, our result implies that spectral clustering can find the hidden clique when its size is at least $c\sqrt{n}$ for some large enough constant c. Our result also provides good insight in understanding the impact of the number of clusters and separation between communities. For instance, in Example 2.2, let $\alpha_n \equiv 1$, $n_{\max} = n_{\min} = n/K$. Then Corollary 3.2 implies that spectral clustering is consistent if $K^2/(n\lambda^2) \to 0$. More generally, the guarantees of Corollary 3.2 compares favorably against most existing results as summarized in Chen, Sanghavi and Xu (2012), in terms of allowable cluster size, density gap and overall sparsity. It would be interesting to develop a unified theoretical framework (e.g., minimax theory) such that all methods and model parameters can be studied and compared together.

APPENDIX: TECHNICAL PROOFS

For any two matrices A and B of the same dimension, we use the notation $\langle A, B\rangle = \operatorname{trace}(A^T B)$ for the standard matrix inner product.

A.1. Proof of Lemma 5.1. By Proposition 2.2 of Vu and Lei (2013), there exists a K-dimensional orthogonal matrix Q such that

$$\frac{1}{\sqrt{2K}}\|\widehat{U} - UQ\|_F \le \frac{1}{\sqrt{K}}\|(I - \widehat{U}\widehat{U}^T)UU^T\|_F \le \|(I - \widehat{U}\widehat{U}^T)UU^T\|.$$

Next, we establish that $\|(I - \widehat{U}\widehat{U}^T)UU^T\| \le 2\frac{\|A-P\|}{\gamma_n}$. If $\|A - P\| \le \gamma_n/2$, then by Davis–Kahan $\sin\Theta$ theorem, we have

$$\|(I - \widehat{U}\widehat{U}^T)UU^T\| \le \frac{\|A - P\|}{\gamma_n - \|A - P\|} \le 2\frac{\|A - P\|}{\gamma_n}.$$

If $\|A - P\| > \gamma_n/2$, then

$$\|(I - \widehat{U}\widehat{U}^T)UU^T\| \le 1 \le 2\frac{\|A - P\|}{\gamma_n}.$$

A.2. Proof of Lemma 5.3. First, by the definition of $\bar{U}$ and the fact that U is feasible for problem (2.2), we have $\|\bar{U} - U\|_F^2 \le 2\|\bar{U} - \widehat{U}\|_F^2 + 2\|\widehat{U} - U\|_F^2 \le (4 + 2\varepsilon)\|\widehat{U} - U\|_F^2$. Then

$$\sum_{k=1}^{K} |S_k|\delta_k^2/4 \le \|\bar{U} - U\|_F^2 \le (4 + 2\varepsilon)\|\widehat{U} - U\|_F^2, \tag{A.1}$$

which concludes the first claim of the lemma.

Under the assumption described in the second part of the lemma, equation (A.1) further implies that

$$|S_k| \le (16 + 8\varepsilon)\|\widehat{U} - U\|_F^2/\delta_k^2 < n_k \qquad \text{for all } k.$$

Therefore, $T_k \equiv G_k \setminus S_k \neq \varnothing$, for each k. If $i \in T_k$ and $j \in T_\ell$ with $k \neq \ell$, then $\bar{U}_{i*} \neq \bar{U}_{j*}$ because otherwise $\max(\delta_k, \delta_\ell) \leq \|U_{i*} - U_{j*}\| \leq \|U_{i*} - \bar{U}_{i*}\| + \|U_{j*} - \bar{U}_{j*}\| < \delta_k/2 + \delta_\ell/2$, which is impossible. This further implies that $\bar{U}$ has exactly K distinct rows, because the number of distinct rows is no larger than K as part of the constraints of the optimization problem (2.2).

On the other hand, if i and j are both in T_k, for some k, then $\bar{U}_{i*} = \bar{U}_{j*}$ because otherwise there would be more than K distinct rows since there are at least $K-1$ other rows occupied by members in T_ℓ for $\ell \neq k$.

As a result, $\bar{U}_{i*} = \bar{U}_{j*}$ if $i, j \in T_k$ for some k, and $\bar{U}_{i*} \neq \bar{U}_{j*}$ if $i \in T_k$, $j \in T_\ell$ with $k \neq \ell$. This gives a correspondence of clustering between the rows in $\bar{U}_{T*}$ and those in U_{T*} where $T = \bigcup_{k=1}^{K} T_k$.

A.3. Proofs for degree corrected block models. The argument fits very well in the general argument developed in Section 5. Then Lemma 5.1 and Theorem 5.2 still apply and

$$\text{(A.2)} \qquad \mathbb{P}\left[\|\widehat{U} - UQ\|_F \leq 2\sqrt{2}C\frac{\sqrt{Kn\alpha_n}}{\gamma_n} \text{ for some } QQ^T = I_K\right] \geq 1 - n^{-1},$$

where C is the constant in Theorem 5.2.

For presentation simplicity, in the following argument we will work with $Q = I_K$. The general case can be handled in the same manner with more complicated notation (simply substitute U by UQ).

To prove Theorem 4.2, we first give a bound on the zero rows in $\widehat{U}$. Recall that $I_+ = \{i : \widehat{U}_{i*} \neq 0\}$. Define $I_0 = I_+^c$.

LEMMA A.1 (Number of zero rows in $\widehat{U}$). *In a DCBM* (Θ, B, ψ) *satisfying the conditions of Theorem* 4.2, *let* $\widehat{U}$ *and* U *be the leading eigenvectors of* A *and* P, *respectively. Then*

$$|I_0| \leq \sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}\, \|\widehat{U} - U\|_F.$$

PROOF. Use Cauchy–Schwarz:

$$\|\widehat{U} - U\|_F^2 \geq \sum_{i=1}^{n} \mathbb{1}(\widehat{U}_{i*} = 0)\|U_{i*}\|^2 \geq \frac{(\sum_{i=1}^{n} \mathbb{1}(\widehat{U}_{i*} = 0))^2}{\sum_{i=1}^{n} \|U_{i*}\|^{-2}} = \frac{|I_0|^2}{\sum_{k=1}^{K} n_k^2 \nu_k}. \qquad \square$$

We also need the following simple fact about the distance between normalized vectors.

FACT. *For two nonzero vectors* v_1, v_2 *of same dimension, we have* $\|\frac{v_1}{\|v_1\|} - \frac{v_2}{\|v_2\|}\| \leq 2\frac{\|v_1 - v_2\|}{\max(\|v_1\|, \|v_2\|)}$.

PROOF. Without loss of generality, assume $\|v_1\| \geq \|v_2\|$. Then

$$\left\|\frac{v_1}{\|v_1\|} - \frac{v_2}{\|v_2\|}\right\| = \left\|\frac{v_1}{\|v_1\|} - \frac{v_2}{\|v_1\|} + \frac{v_2}{\|v_1\|} - \frac{v_2}{\|v_2\|}\right\|$$
$$\leq \frac{\|v_1 - v_2\|}{\|v_1\|} + \frac{\|v_2\| \big| \|v_1\| - \|v_2\| \big|}{\|v_1\|\|v_2\|} \leq 2\frac{\|v_1 - v_2\|}{\|v_1\|}. \qquad \square$$

PROOF OF THEOREM 4.2. Recall that U' is the row-normalized version of U. Let $U'' = U'_{I_+ *}$ be the sub-matrix of U' corresponding to the nonzero rows in $\widehat{U}$. Then

$$\|\widehat{U}' - U''\|_{2,1} \leq 2\sum_{i=1}^{n} \frac{\|\widehat{U}_{i*} - U_{i*}\|}{\|U_{i*}\|}$$
$$\leq 2\sqrt{\sum_{i=1}^{n} \|\widehat{U}_{i*} - U_{i*}\|^2 \sum_{i=1}^{n} \|U_{i*}\|^{-2}} \leq 2\sqrt{\|\widehat{U} - U\|_F^2 \sum_{k=1}^{K} n_k^2 \nu_k}.$$

Now we can bound the (2, 1) distance between an approximate solution of k-median problem (4.2) and the targeted solution U''.

$$\|\widehat{\Theta}^+\widehat{X} - U''\|_{2,1} \leq \|\widehat{\Theta}^+\widehat{X} - \widehat{U}'\|_{2,1} + \|\widehat{U}' - U''\|_{2,1}$$
$$\leq (2+\varepsilon)\|\widehat{U}' - U''\|_{2,1}.$$

Let $S = \{i \in I_+ : \|\widehat{\Theta}_{i*}\widehat{X} - U'_{i*}\| \geq \frac{1}{\sqrt{2}}\}$. The size of S can be bounded using a similar argument as in the proof of Lemma A.1.

$$|S|\frac{1}{\sqrt{2}} \leq \|\widehat{\Theta}^+\widehat{X} - U''\|_{2,1} \leq (2+\varepsilon)\|\widehat{U}' - U''\|_{2,1}$$
$$\leq 2(2+\varepsilon)\sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}\|\widehat{U} - U\|_F,$$

which implies

$$|S| \leq 2\sqrt{2}(2+\varepsilon)\sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}\|\widehat{U} - U\|_F. \tag{A.3}$$

On the event in (A.2) (recall that we assume $Q = I$), (A.3) and Lemma A.1 implies

$$|S| + |I_0| \leq (2.5+\varepsilon)8C\frac{\sqrt{Kn\alpha_n}}{\gamma_n}\sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}. \tag{A.4}$$

Combining this with condition (4.3) implies $|S| + |I_0| < n_k$ for all k and hence $G_k \cap (I_+ \setminus S) \neq \varnothing$. Therefore, for any two rows in $G := I_+ \setminus S$, if they are in different clusters of Θ then they must be in different clusters of $\widehat{\Theta}$ (otherwise, $\|U'_{i*} - U'_{j*}\| \leq \|U'_{i*} - \widehat{\Theta}_{i*}\widehat{X}\| + \|\widehat{\Theta}_{j*}\widehat{X} - U'_{j*}\| < \sqrt{2}$).

As a consequence, the mis-clustered nodes are no more than $I_0 \cup S$, and the number is bounded by the right-hand side of (A.4). The claimed result follows by choosing $c = 8C$. □

Acknowledgment. The authors thank an anonymous reviewer for helpful suggestions that led in particular to a significant simplification of the proof of Lemma 5.1.

SUPPLEMENTARY MATERIAL

Supplement to "Consistency of spectral clustering in sparse stochastic block models" (DOI: 10.1214/14-AOS1274SUPP; .pdf). The supplementary file contains a proof of Theorem 5.2.

REFERENCES

ALOISE, D., DESHPANDE, A., HANSEN, P. and POPAT, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning* **75** 245–248.

ALON, N. and SPENCER, J. H. (2004). *The Probabilistic Method*, 2nd ed. Wiley, Hoboken.

AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2012). Pseudo-likelihood methods for community detection in large sparse networks. Preprint. Available at arXiv:1207.2340.

ANANDKUMAR, A., GE, R., HSU, D. and KAKADE, S. M. (2013). A tensor spectral approach to learning mixed membership community models. Preprint. Available at arXiv:1302.2684.

AWASTHI, P. and SHEFFET, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Lecture Notes in Computer Science* **7408** 37–49. Springer, Heidelberg. MR3003539

BALAKRISHNAN, S., XU, M., KRISHNAMURTHY, A. and SINGH, A. (2011). Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems* 24 (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, eds.) 954–962. Curran Associates, Red Hook, NY.

BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467

CHANNAROND, A., DAUDIN, J.-J. and ROBIN, S. (2012). Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electron. J. Stat.* **6** 2574–2601. MR3020277

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* 1–10. ACM, New York, NY.

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 (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 2204–2212. Curran Associates, Red Hook, NY.

CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253

CHUNG, F. and RADCLIFFE, M. (2011). On the spectra of general random graphs. *Electron. J. Combin.* **18** Paper 215, 14. MR2853072

COJA-OGHLAN, A. (2010). Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** 227–284. MR2593622

DECELLE, A., KRZAKALA, F., MOORE, C. and ZDEBOROVÁ, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* (3) **84** 066106.

DESHPANDE, Y. and MONTANARI, A. (2013). Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. Preprint. Available at arXiv:1304.7047.

FEIGE, U. and OFEK, E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures Algorithms* **27** 251–275. MR2155709

FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning* **2** 129–233.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

JIN, J. (2012). Fast community detection by SCORE. Available at arXiv:1211.5803.

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206

KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer, New York. MR2724362

KRZAKALA, F., MOORE, C., MOSSEL, E., NEEMAN, J., SLY, A., ZDEBOROVÁ, L. and ZHANG, P. (2013). Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. USA* **110** 20935–20940. MR3174850

KUMAR, A. and KANNAN, R. (2010). Clustering with spectral norm and the k-means algorithm. In *Proceedings of the* 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science FOCS* 299–308. IEEE, Los Alamitos, CA. MR3025203

KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proceedings of the* 45*th Annual IEEE Symposium on Foundations of Computer Science* 454–462. IEEE Computer Society, Washington, DC.

LEI, J. and RINALDO, A. (2014). Supplement to "Consistency of spectral clustering in stochastic block models." DOI:10.1214/14-AOS1274SUPP.

LI, S. and SVENSSON, O. (2013). Approximating k-median via pseudo-approximation. In *Proceedings of the* 45*th Annual ACM Symposium on Symposium on Theory of Computing* 901–910. ACM, New York.

LU, L. and PENG, X. (2012). Spectra of edge-independent random graphs. Preprint. Available at arXiv:1204.6207.

LYZINSKI, V., SUSSMAN, D., TANG, M., ATHREYA, A. and PRIEBE, C. (2013). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. Preprint. Available at arXiv:1310.0532.

MASSOULIE, L. (2013). Community detection thresholds and the weak Ramanujan property. Preprint. Available at arXiv:1311.3085.

MCSHERRY, F. (2001). Spectral partitioning of random graphs. In *42nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. IEEE, Los Alamitos, CA. MR1948742

MOSSEL, E., NEEMAN, J. and SLY, A. (2012). Stochastic block models and reconstruction. Preprint. Available at arXiv:1202.1499.

MOSSEL, E., NEEMAN, J. and SLY, A. (2013). A proof of the block model threshold conjecture. Preprint. Available at arXiv:1311.4115.

NEWMAN, M. E. J. (2010). *Networks*: *An Introduction*. Oxford Univ. Press, Oxford. MR2676073

NEWMAN, M. E. J. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* (3) **69** 026113.

NG, A. Y., JORDAN, M. I., WEISS, Y. et al. (2002). On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2** 849–856.

QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. Preprint. Available at arXiv:1309.4111.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

SARKAR, P. and BICKEL, P. (2013). Role of normalization in spectral clustering for stochastic blockmodels. Preprint. Available at arXiv:1310.1495.

SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. MR3010899

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803

VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
132 BAKER HALL/5000 FORBES AVE.
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jinglei@andrew.cmu.edu
arinaldo@cmu.edu