Consistency of Spectral Clustering in Stochastic Block Models

Jing Lei\textsuperscript{1} and Alessandro Rinaldo\textsuperscript{2}

\textit{Carnegie Mellon University}

September 23, 2014

Abstract

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 (2009), and Goldenberg et al. (2010).

Among the many existing statistical models for network data, the stochastic block model, henceforth SBM, of Holland et al. (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 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 & Girvan, 2004), likelihood methods (Bickel & Chen, 2009; Choi et al., 2012; Zhao et al.,

\textsuperscript{1}Jing Lei’s research was partially supported by NSF Grant BCS-0941518, NSF Grant DMS-1407771 and NIH Grant MH057881.

\textsuperscript{2}Alessandro Rinaldo’s research was partially supported by the grant FA9550-12-1-0392 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) and by the NSF CAREER grant DMS-1149677.
2012; Amini et al., 2012; Celisse et al., 2012), method of moments (Anandkumar et al., 2013), belief propagation (Decelle et al., 2011), convex optimization (Chen et al., 2012), spectral clustering (Rohe et al., 2011; Balakrishnan et al., 2011; Jin, 2012; Fishkind et al., 2013; Sarkar & Bickel, 2013) and its variants (Coja-Oghlan, 2010; Chaudhuri et al., 2012) and spectral embeddings (Sussman et al., 2012, 2013; 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, e.g., Rohe et al. (2011); Jin (2012); Sarkar & 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, Rohe et al. (2011); Fishkind et al. (2013) require the expected node degree to be almost linear in \( n \), while Jin (2012) requires polynomial growth. Analogous conditions can be found elsewhere: see, e.g., Sussman et al. (2013), Balakrishnan et al. (2011), and Sussman et al. (2012).

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 & 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 & 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 McSherry (2001); Chen et al. (2012); Channarond et al. (2012); Sarkar & 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; Mossel et al., 2012; Krzakala et al., 2013; Massoulie, 2013; Mossel et al., 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 memberships 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 et al. (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 & 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 $I, J \subseteq [n]$, let $M_{I*}$ and $M_{*,J}$ be the submatrix of $M$ consisting the corresponding rows and columns. Let $\mathcal{M}_{n,K}$ be the collection of all $n \times K$ matrices where each row has exactly one 1 and $(K - 1)$ 0’s. For any $\Theta \in \mathcal{M}_{n,K}$, we call $\Theta$ a membership matrix, and the community membership of a node $i$ is denoted by $g_i \in \{1, ..., K\}$, which satisfies $\Theta_{ig_i} = 1$. Let $G_k = G_k(\Theta) = \{1 \leq i \leq n : g_i = k\}$ and $n_k = |G_k|$ for all $1 \leq k \leq K$. Let $n_{\min} = \min_{1 \leq k \leq K} n_k$, $n_{\max} = \max_{1 \leq k \leq K} n_k$, 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. $\|M\|_F = (\text{trace}(MM^T))^{1/2}$ denotes the Frobenius norm of a matrix $M$. The $\ell_0$ norm $\|M\|_0$ simply counts the number of non-zero 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 $|x_n/y_n| \leq C$ 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 $(\Theta, B)$, where $\Theta \in \mathcal{M}_{n,K}$ 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 $\leq g_i \leq K$) be its community label, such that the $i$th row of $\Theta$ 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 $\ell$. Given $(\Theta, B)$, the adjacency
matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ is generated as

$$a_{ij} = \begin{cases} \text{independent Bernoulli}(B_{g_{ij}}) & \text{if } i < j, \\ 0 & \text{if } i = j, \\ a_{ji} & \text{if } i > j. \end{cases}$$

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

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

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

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

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

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

### 2.2 Spectral clustering

Spectral clustering is a simple method for community recovery (von Luxburg, 2007; Rohe et al., 2011; Jin, 2012). 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 $\mathbb{E}(A) = P - \text{diag}(P)$. Let $P = UD_1U^T$ be the eigen-decomposition of $P$ with $U^TU = I_K$ 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 $(\Theta, B)$ parametrize a SBM with $K$ communities, where $B$ is full rank. Let $UD_1U^T$ be the eigen-decomposition of $P = \Theta B\Theta^T$. Then $U = \Theta X$ where $X \in \mathbb{R}^{K \times K}$ and $\|X_{k\ell} - X_{\ell k}\| = \sqrt{n_k^{-1} + n_{\ell}^{-1}}$ for all $1 \leq k \leq K$.

**Proof.** Let $\Delta = \text{diag}(\sqrt{n_1}, \ldots, \sqrt{n_K})$ then

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

It is straightforward to verify that $\Theta \Delta^{-1}$ is orthonormal. Let $Z\Delta^T = DB\Delta$ be the eigen-decomposition of $DB\Delta$. 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 $k$th row has length $\|((\Delta Z)_{k\ell})\|=\sqrt{1/n_k}$.

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 $P$:

$$A - P = (A - \mathbb{E}(A)) - \text{diag}(P),$$
Algorithm 1: Spectral Clustering with Approximate \( k \)-means

**Input:** Adjacency matrix \( A \); number of communities \( K \); approximation parameter \( \epsilon \).

**Output:** Membership matrix \( \hat{\Theta} \in \mathbb{M}_{n,K} \).

1. Calculate \( \hat{U} \in \mathbb{R}^{n \times K} \) consisting of the leading \( k \) eigenvectors (ordered in absolute eigenvalue) of \( A \).
2. Let \((\hat{\Theta}, \hat{X})\) be an \((1 + \epsilon)\)-approximate solution to the \( k \)-means problem (3) with \( K \) clusters and input matrix \( \hat{U} \).
3. Output \( \hat{\Theta} \).

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 with \( n \) while the noise matrix \((A - \mathbb{E}(A))\) has operator norm on the scale of \( \sqrt{n} \) and \( \text{diag}(P) \) is like a constant. Therefore, letting \( A = \hat{U} \hat{D} \hat{U}^T \) be the \( K \)-dimensional eigen-decomposition of \( A \) corresponding to the \( K \) largest absolute eigenvalues, we can see that \( \hat{U} \) 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 \( \hat{U} \). In this paper we consider the \( k \)-means clustering, defined as

\[
(\hat{\Theta}, \hat{X}) = \arg \min_{\Theta \in \mathbb{M}_{n,K}, X \in \mathbb{R}^{K \times K}} \| \Theta X - \hat{U} \|_F^2.
\]  

(2)

It is known that finding a global minimizer for the \( k \)-means problem (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 et al., 2004). That is, there are polynomial time algorithms that find

\[
(\hat{\Theta}, \hat{X}) \in \mathbb{M}_{n,K} \times \mathbb{R}^{K \times K}, \\
\text{s.t.} \quad \| \hat{\Theta} \hat{X} - \hat{U} \|_F^2 \leq (1 + \epsilon) \min_{\Theta \in \mathbb{M}_{n,K}, X \in \mathbb{R}^{K \times K}} \| \Theta X - \hat{U} \|_F^2.
\]  

(3)

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 \((\Theta, B)\) where

\[
B = \alpha_n B_0; \quad B_0 = \lambda I_K + (1 - \lambda)1_K 1_K^T, \quad 0 < \lambda < 1,
\]  

(4)

\( I_K \) is the \( K \times K \) identity matrix, and \( 1_K \) 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 $\alpha_n$ within community and $\alpha_n(1 - \lambda)$ between community. The quantity $\lambda$ reflects the relative difference in connectivity between communities and within communities. The network sparsity is captured by $\alpha_n$, 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 $\alpha_n$ or $\lambda$ 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

$$B = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$  \hspace{1cm} (5)

In the planted clique problem, it is known that community recovery is easy if $n_1 \geq c\sqrt{n}$ for a constant $c$ (see Deshpande & 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 $\alpha_n$ on the performance of spectral clustering. Nevertheless, our results explicitly keep track of other quantities such as $K$, $\lambda$, $n_{\max}$, and $n_{\min}$, all of which are allowed to change with $n$ in a non-trivial manner. The dependence of recovery error bound on some of these quantities, such as $K$ and $\lambda$, is concerned by some authors, such as Chen et al. (2012); Chaudhuri et al. (2012); Anandkumar et al. (2013). For ease of readability, we do not always make this dependence on $n$ explicit in our notation.

### 3 Main results for stochastic block model

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

**Theorem 3.1.** Let $A$ be an adjacency matrix generated from a stochastic block model $(\Theta, B)$. Assume that $P = \Theta B \Theta^T$ is of rank $K$, with smallest absolute non-zero eigenvalue at least $\gamma_n$ and $\max_{k,\ell} B_{k\ell} \leq \alpha_n$ for some $\alpha_n \geq \log n/n$. Let $\hat{\Theta}$ be the output of spectral clustering using $(1 + \epsilon)$ approximate $k$-means (Algorithm 1). There exists an absolute constant $c > 0$, such that, if

$$(2 + \epsilon) \frac{K n \alpha_n}{\gamma_n^2} < c,$$  \hspace{1cm} (6)

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

$$\sum_{k=1}^n \frac{|S_k|}{n_k} \leq c^{-1} (2 + \epsilon) \frac{K n \alpha_n}{\gamma_n^2}.$$  \hspace{1cm} (7)

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

The sets $S_k$ ($1 \leq k \leq K$) 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 $\tilde{L}(\hat{\Theta}, \Theta)$ (and hence on $L(\hat{\Theta}, \Theta)$) through eq. (7).
Condition (6) specifies the range of model parameters \((K, n, \gamma_n, \alpha_n)\) for which the result is applicable. It is included only for technical reasons, because it holds whenever the bound in (7) vanishes and therefore implies consistency. In particular, as discussed after Corollary 3.2, we have \(K n \alpha_n/\gamma_n^2 = o(1)\) in many interesting cases. The constant \(c\) in (6) 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 Appendix B. The assumption of \(\alpha_n \geq \log n / n\) can be changed to \(\alpha_n \geq c_0 \log n / n\) for any \(c_0 > 0\), and also the probability bound \(1 - n^{-r}\) can be changed to \(1 - n^{-r}\) for any \(r > 0\), with a different constant \(c = c(c_0, r)\) in (6) and (7).

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 \(\alpha_n\) decreases. This is because the quantity \(\gamma_n\) contains the parameter \(\alpha_n\). Also the dependence on the community size imbalance as well as the community separation (which corresponds to the parameter \(\lambda\) 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 \((\Theta, B)\), where \(B = \alpha_n B_0\) for some \(\alpha_n \geq \log n / n\) and with \(B_0\) having minimum absolute eigenvalue \(\geq \lambda > 0\) and \(\max_k \|B_0(k, \ell)\| = 1\). Let \(\Theta\) be the output of spectral clustering using \((1 + \epsilon)\) approximate \(k\)-means (Algorithm 1). Then, there exists an absolute constant \(c\) such that if

\[
(2 + \epsilon) \frac{Kn}{n^2_{\min} \lambda^2 \alpha_n} < c
\]

then with probability at least \(1 - n^{-1}\),

\[
\tilde{L}(\Theta, \Theta) \leq c^{-1}(2 + \epsilon) \frac{Kn}{n^2_{\min} \lambda^2 \alpha_n},
\]

and

\[
L(\Theta, \Theta) \leq c^{-1}(2 + \epsilon) \frac{Kn'_{\max}}{n^2_{\min} \lambda^2 \alpha_n}.
\]

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

In Example 2.2, the smallest non-zero eigenvalue of \(B_0\) is \(\lambda\). Recall that \(\lambda\) 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 \(\alpha_n\) 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, \(L(\Theta, \Theta)\) has limited meaning because a trivial clustering putting all nodes in one cluster achieves \(L(\Theta, \Theta) = 2n_{\min} / n\) which is \(o(1)\) in the most interesting regime. Therefore, it makes more sense to consider \(\tilde{L}(\Theta, \Theta)\). Now \(B_0 = B\) is given by (5), with minimum eigenvalue \(> 0.19\). Applying Corollary 3.2 with \(K = 2\), \(\lambda = 0.19\), \(\alpha_n = 1\), and any fixed \(\epsilon > 0\), we have

\[
\tilde{L}(\Theta, \Theta) < c' \frac{n}{n_{\min}^2}.
\]
provided that \(c'n/n^2_{\text{min}} < 1\), where \(c'\) is a different absolute constant. Therefore, when \(n_{\text{min}} \geq \sqrt{an}\) for some \(a > c'\), \(\hat{\Theta}\) 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 & 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,

\[
\lambda^2 \alpha_n^2 n > c \sigma_n^2 \log n, \quad \text{and} \quad \sigma_n^2 > (\log n)^6/n, \tag{9}
\]

for some constant \(c\), where \(\sigma_n^2\) is an upper bound on the maximal variance of the edges. Therefore, the condition (9) implies that \(\alpha_n > \sqrt{\lambda^{-1} (\log n)^{3.5}}/n\), which is stronger than the condition in our Corollary 3.2.

4 Degree corrected block models

The degree corrected block model (DCBM, Karrer & 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 \((\Theta, B, \psi)\), where, in addition to the membership matrix \(\Theta\) 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 \((\Theta, B, \psi)\), the edge probability between nodes \(i\) and \(j\) is \(\psi_i \psi_j B_{ij}g_ij\) (recall that \(g_i\) is the community label of node \(i\)). Similar to the SBM, the DCBM also assumes independent edge formation given \((\Theta, B, \psi)\). The inclusion of \(\psi\) raises an issue of identifiability. So we assume that \(\max_{i \in G_k} \psi_i = 1\) for all \(k = 1, ..., K\). 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 et al. (2012) for likelihood methods, and in Chaudhuri et al. (2012); Jin (2012) for spectral methods.

Additional notation about the degree heterogeneity

Let \(\phi_k\) be the \(n \times 1\) vector that agrees with \(\psi\) 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 \(\tilde{n}_{\text{min}} = \min_k \tilde{n}_k\) and \(\tilde{n}_{\text{max}} = \max_k \tilde{n}_k\).

The spectral clustering heuristic can be extended to DCBM’s by considering the eigen-decomposition \(P = UDU^T\) where \(P = \text{diag}(\psi) \Theta B \Theta^T \text{diag}(\psi)\). Now the matrix \(U\) may have more than \(K\) distinct rows due to the effect of \(\psi\). 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 DCBM’s.

Lemma 4.1 (Spectral structure of mean matrix in DCBM). Let \(UDU^T\) be the eigen-decomposition of \(P = \text{diag}(\psi) \Theta B \Theta^T \text{diag}(\psi)\) in a DCBM parameterized by \((\Theta, B, \psi)\). Then there exists a \(K \times K\)
orthogonal matrix $H$ such that
\[ U_{is} = \tilde{\psi}_i H_{ks}, \quad \forall \ 1 \leq k \leq K, \ i \in G_k. \]

**Proof.** First realize that $\text{diag}(\psi) \Theta = \tilde{\Theta} \Psi$, where $\Psi = \text{diag}(\|\phi_1\|, \ldots, \|\phi_K\|)$.

\[
P = \text{diag}(\psi) \Theta B \Theta^T \text{diag}(\psi) = \tilde{\Theta} \Psi \tilde{\Theta}^T = \tilde{\Theta} H D (\tilde{\Theta} H)^T
\]

(10)

where $\Psi B \Psi = HDH^T$ is the eigen-decomposition of $\Psi B \Psi$. Note that $\tilde{\Theta}^T \tilde{\Theta} = I_K$ so $\tilde{\Theta} HD (\tilde{\Theta} H)^T$ 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 $\psi$. In particular, we consider spherical clustering, which looks for a cluster structure among the rows of a normalized matrix $U'$ with $U'_{is} = U_{is}/\|U_{is}\|$.

In addition to the overall sparsity, the difficulty of community recovery in a DCBM is also affected by small entries of $\psi$. 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 $\psi$ and the overall network sparsity (the maximum/average degree) has not been well understood. In the analysis of profile likelihood methods, Zhao et al. (2012) assume that the entries of $\psi$ are fixed constants. In spectral clustering, Jin (2012) allows milder conditions on $\psi$ 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 $G_k$:

\[
\nu_k := n_k^{-2} \sum_{i \in G_k} \tilde{\psi}_i^{-2}, \quad k = 1, 2, \ldots, K.
\]

By definition $\nu_k \in [1, \infty)$ and a larger $\nu_k$ indicates a stronger heterogeneity in the $k$th 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 SBM’s in previous sections can be extended to cover very general degree corrected models. In particular, let $\hat{U} \in \mathbb{R}^{n \times K}$ consist the $K$ leading eigenvectors of $A$. We consider the following spherical $k$-median spectral clustering:

\[
\text{minimize}_{\Theta \in \mathbb{M}_{n,K}, X \in \mathbb{R}^{K \times K}} \|\Theta X - \hat{U}'\|_{2,1},
\]

(11)

where $\hat{U}'$ is the row-normalized version of $\hat{U}$ and $\|M\|_{2,1} = \sum_{i=1}^n \|M_{i*}\|$ is the matrix $(2,1)$-norm. We will not require to solve (11) exactly but instead we consider a $(1 + \epsilon)$ approximation $(\hat{\Theta}, \hat{X})$ to the $k$-median problem, which can be solved in polynomial time when $\epsilon > \sqrt{3}$ (Charikar et al., 1999; Li & Svensson, 2013). The practical procedure will also take care of the possible zero rows in $\hat{U}$ 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 DCBM’s. It is proved in Appendix A.3.

**Theorem 4.2** (Main result for DCBM). Consider a DCBM $(\Theta, B, \psi)$ with $K$ communities, where $P = \text{diag}(\psi) \Theta B \Theta^T \text{diag}(\psi)$ has rank $K$, the smallest non-zero absolute eigenvalue at least $\gamma_n$, and
Algorithm 2: Spherical $k$-median Spectral Clustering

**Input:** Adjacency matrix $A$; number of communities $K$; approximation parameter $\epsilon$.

**Output:** Membership matrix $\hat{\Theta} \in \mathbb{M}_{n,K}$.

1. Calculate $\hat{U} \in \mathbb{R}^{n \times K}$ consisting of the leading $k$ eigenvectors (ordered in absolute eigenvalue) of $A$.
2. Let $I_+ = \{ i : \| \hat{U}_{i*} \| > 0 \}$ and $\hat{U}^+ = (\hat{U}_{I_+})$.
3. Let $\hat{U}'$ be row-normalized version of $\hat{U}^+$.
4. Let $(\hat{\Theta}^+, \hat{X})$ be an $(1+\epsilon)$-approximate solution to the $k$-median problem with $K$ clusters and input matrix $\hat{U}'$.
5. Output $\hat{\Theta}$ with $\hat{\Theta}_{i*}$ being the corresponding row in $\hat{\Theta}^+$ if $i \in I_+$, and $\hat{\Theta}_{i*} = (1, 0, \ldots, 0)$ if $i \notin I_+$.

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

$$
(2.5 + \epsilon) \frac{\sqrt{Kn\alpha_n}}{\gamma_n} < c \frac{n_{\min}}{\sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}} \quad (12)
$$

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

$$
L(\hat{\Theta}, \Theta) \leq c^{-1}(2.5 + \epsilon) \sqrt{\sum_{k=1}^{K} n_k^2 \nu_k} \frac{\sqrt{Kn\alpha_n}}{\gamma_n \sqrt{n}}. \quad (13)
$$

**Remark.** The constant $c$ equals $1/(8C)$ where $C$ is the universal constant in Theorem 5.2. The condition on $\alpha_n$ 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 eqs. (12) and (13).

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 $(\Theta, B, \psi)$, 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 $(\hat{\Theta}, \hat{X})$ be an $(1 + \epsilon)$–approximate solution to the spherical $k$-median algorithm (Algorithm 2). There exists an absolute constant $c$ such that if

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

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

$$
L(\hat{\Theta}, \Theta) \leq c^{-1}(2.5 + \epsilon) \frac{\sqrt{K}}{n_{\min} \lambda^{\sqrt{\alpha_n}}} \sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}. \quad (14)
$$
Comparing with Theorem 3.1 and Corollary 3.2, the results for DCBM are different in two major aspects. First, the DCBM condition (12) 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 (12) more stringent than (6). Also the upper bound on $L(\hat{\Theta}, \Theta)$ 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 $\tilde{L}(\hat{\Theta}, \Theta)$. 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 $\psi$ has comparable degree heterogeneity across communities: $c_1 \nu \leq \nu_k \leq c_2 \nu$ for constants $c_1, c_2$. Then Corollary 4.3 implies an overall relative error rate

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

(14)

Several observations are worth mentioning. First, the error rate depends on $\nu$, 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 $\psi_i$’s bounded away from zero (but the rest should not be too small in order to keep $\nu$ small). Third, if there is no degree heterogeneity ($\nu_k \equiv 1$ and $\tilde{n}_{\min} = n_{\min}$), then the rate in (14) 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 $\alpha_n$ is as small as $\log n/n$, provided that $1/\nu$, $\tilde{n}_{\min}/n$, and $\lambda$ 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 et al. (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 $\alpha_n$ 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 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 non-zero singular value $\gamma_n$. Let $A$ be any symmetric matrix and $\bar{U}, U \in \mathbb{R}^{n \times K}$ be the $K$ leading eigenvectors of $A$ and $P$ respectively. Then there exists a $K \times K$ orthogonal matrix $Q$ such that

$$\|\bar{U} - UQ\|_F \leq 2\sqrt{2K}\gamma_n\|A - P\|.$$

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 $E(a_{ij}) = p_{ij}$.

**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 $E[A] = P = (p_{ij})_{i,j=1,...,n}$ and assume that $n \max_{i,j} p_{ij} \leq d$ for $d \geq c_0 \log n$ and $c_0 > 0$. Then, for any $r > 0$ there exists a constant $C = C(r, c_0)$ such that

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

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 & Peng (2012) use a path counting technique in random matrix theory to prove a bound of the same order but require a maximal degree $d \geq c_0 (\log n)^4$.

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 & Ofek (2005) for bounding the second largest eigenvalue of an Erdős-Rényi random graph with edge probability $d/n$. The full proof is provided in Appendix B. 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 & 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 $\hat{U}$ and the ideal input matrix $U$.

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

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

$$\sum_{k=1}^{K} |S_k| \delta_k^2 \leq 4(4 + 2\epsilon)\|\hat{U} - U\|_F^2. \quad (15)$$

Moreover, if

$$(16 + 8\epsilon)\|\hat{U} - U\|_F^2 / \delta_k^2 < n_k \quad \text{for all } k, \quad (16)$$

then there exists a $K \times K$ permutation matrix $J$ such that $\hat{\Theta}_{G*} = \Theta_{G*} J$, 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 & Sheffet (2012) (see also Kumar & 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 $Q$,

$$\|\hat{U} - UQ\|_F \leq \frac{2\sqrt{2K}}{\gamma_n} \|A - P\| \leq \frac{2\sqrt{2K}}{\gamma_n} C \sqrt{n \alpha_n}, \quad (17)$$

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 $\hat{U}$ and $UQ$. To that end, Lemma 2.1 implies that $UQ = \Theta XQ = \Theta X'$ where $\|X'_{k*} - X'_{\ell*}\| = \sqrt{\frac{1}{n_k} + \frac{1}{n_\ell}}$. As a result, we can choose $\delta_k = \sqrt{\frac{1}{n_k} + \frac{1}{\max(n_{\ell}: \ell \neq k)}}$ in Lemma 5.3 and hence $n_k \delta_k^2 \geq 1$ for all $k$. Using (17), a sufficient condition for (16) to hold is

$$(16 + 8\epsilon)8C^2K \frac{n \alpha_n}{\gamma_n^2} \leq 1 \leq \min_{1 \leq k \leq K} n_k \delta_k^2, \quad (18)$$
so that (6) indeed implies (16) by setting $c = \frac{1}{64C_2}$. In detail, the choice of $\delta_k = 1/\sqrt{n_k}$ together with (15) yields that

$$\sum_{k=1}^{K} |S_k| \left( \frac{1}{n_k} + \frac{1}{\max\{n_{\ell} : \ell \neq k\}} \right) = \sum_{k=1}^{K} |S_k| \delta_k^2 \leq 4(4 + 2\epsilon)\|\hat{U} - UQ\|_F^2.$$ 

which, combined with (17), gives (7):

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

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

**Proof of Corollary 3.2.** It is easy to see, e.g., from (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

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

which implies that

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

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

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

## 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 (von Luxburg, 2007; Rohe et al., 2011; Sarkar & Bickel, 2013). 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 et al. (2012); Qin & 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 $\psi$. The intuition is that $\nu_k$ can be very large even when only one $\psi_i$ 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 ($K$, $\alpha_n$, $\lambda$, $n_{\text{min}}$, $n_{\text{max}}$). Existing results usually develop error bounds depending on a subset of these parameters, keeping others as constant (see Bickel & Chen, 2009; Chen et al., 2012; Zhao et al., 2012, for example). 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 $\alpha_n \equiv 1$, $n_{\text{max}} = n_{\text{min}} = n/K$. 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 et al. (2012), in terms of allowable cluster size, density gap and overall sparsity. It would be interesting to develop a unified theoretical framework (for example, minimax theory) such that all methods and model parameters can be studied and compared together.

Acknowledgments

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

A Technical Proofs

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

A.1 Proof of Lemma 5.1

By proposition 2.2 of Vu & Lei (2013), there exists a $K$ dimensional orthogonal matrix $Q$ such that

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

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

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

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

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

A.2 Proof of Lemma 5.3

First by the definition of $\hat{U}$ and the fact that $U$ is feasible for problem (2) we have $\|\hat{U} - U\|_F^2 \leq 2\|\hat{U} - \hat{U}\|_F^2 + 2\|\hat{U} - U\|_F^2 \leq (4 + 2\varepsilon)\|\hat{U} - U\|_F^2$. Then

$$\sum_{k=1}^K |S_k|^2/4 \leq \|\hat{U} - U\|_F^2 \leq (4 + 2\varepsilon)\|\hat{U} - U\|_F^2. \quad (19)$$
which concludes the first claim of the lemma.

Under the assumption described in the second part of the Lemma, Equation (19) further implies that

$$|S_k| \leq (16 + 8e)\|U - U\|^2_F/\delta_k^2 < n_k, \text{ for all } k.$$ 

Therefore $T_k \equiv G_k \setminus S_k \neq \emptyset$, for each $k$. If $i \in T_k$ and $j \in T_{\ell}$ with $k \neq \ell$, then $\bar{U}_{is} \neq \bar{U}_{js}$ because otherwise $\max(\delta_k, \delta_{\ell}) \leq \|U_{is} - U_{js}\| \leq \|U_{is} - \bar{U}_{is}\| + \|U_{js} - \bar{U}_{js}\| < \delta_k/2 + \delta_{\ell}/2$, 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).

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

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

### 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

$$\mathbb{P}\left[\|\bar{U} - UQ\|_F \leq 2\sqrt{2}C\sqrt{\frac{Kna}{\gamma_n}} \quad \text{for some } QQ^T = I_K\right] \geq 1 - n^{-1}, \quad (20)$$

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 $\bar{U}$. Recall that $I_+ = \{i : \bar{U}_{is} \neq 0\}$. Define $I_0 = I_+^c$.

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

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

**Proof.** Use Cauchy-Schwartz:

$$\|\bar{U} - U\|^2_F \geq \sum_{i=1}^n I(\bar{U}_{is} = 0)\|U_{is}\|^2 \geq \left(\sum_{i=1}^n I(\bar{U}_{is} = 0)\right)^2 \sum_{i=1}^n \|U_{is}\|^{-2} = \frac{|I_0|^2}{\sum_{k=1}^K n_k^2 \nu_k}.$$

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

**Fact.** *For two non-zero vectors $v_1, v_2$ of same dimension we have $\|v_1 - v_2\|/\|v_1\| \leq 2\max(\|v_1\|, \|v_2\|)$.***
Proof. Without loss of generality, assume \( \|v_1\| \geq \|v_2\| \). Then
\[
\left\| \frac{v_1}{\|v_1\|} - \frac{v_2}{\|v_2\|} \right\| = \left\| \frac{v_1}{\|v_1\|} - \frac{v_2}{\|v_1\|} + \frac{v_2}{\|v_1\|} - \frac{v_2}{\|v_2\|} \right\| \\
\leq \left\| \frac{v_1 - v_2}{\|v_1\|} \right\| + \left\| \frac{v_2}{\|v_1\|} - \frac{v_2}{\|v_2\|} \right\| \leq 2 \left\| \frac{v_1 - v_2}{\|v_1\|} \right\| .
\]
\[\square\]

Proof of Theorem 4.2. Recall that \( U' \) is the row-normalized version of \( U \). Let \( U'' = U'_{I + \cdot} \) be the submatrix of \( U' \) corresponding to the non-zero rows in \( \hat{U} \). Then
\[
\|\hat{U}' - U''\|_{2,1} \leq 2 \sum_{i=1}^{n} \frac{\|\hat{U}_{i \cdot} - U_{i \cdot}\|}{\|U_{i \cdot}\|} \\
\leq 2 \left( \sum_{i=1}^{n} \|\hat{U}_{i \cdot} - U_{i \cdot}\|^2 \sum_{i=1}^{n} \|U_{i \cdot}\|^2 \right)^{-1/2} \leq 2 \sqrt{\|\hat{U}' - U''\|_F^2 \sum_{k=1}^{K} n_k^2 \nu_k}. 
\]

Now we can bound the \((2,1)\) distance between an approximate solution of \( k \)-median problem (11) and the targeted solution \( U'' \).
\[
\|\hat{X}^+ - U''\|_{2,1} \leq \|\hat{X}^+ - \hat{U}'\|_{2,1} + \|\hat{U}' - U''\|_{2,1} \\
\leq (2 + \epsilon)\|\hat{U}' - U''\|_{2,1}.
\]

Let \( S = \{i \in I_+ : \|\hat{X}_{i \cdot} - U'_{i \cdot}\| \geq \frac{1}{\sqrt{2}}\} \). The size of \( S \) can be bounded using a similar argument as in the proof of Lemma A.1.
\[
|S| \frac{1}{\sqrt{2}} \leq \|\hat{X}^+ - U''\|_{2,1} \leq (2 + \epsilon)\|\hat{U}' - U''\|_{2,1} \\
\leq 2(2 + \epsilon) \sum_{k=1}^{K} n_k^2 \nu_k \|\hat{U}' - U''\|_F,
\]
which implies
\[
|S| \leq 2\sqrt{2}(2 + \epsilon) \sqrt{\sum_{k=1}^{K} n_k^2 \nu_k \|\hat{U}' - U''\|_F}.
\]

On the event in (20) (recall that we assume \( Q = I \)), (21) and Lemma A.1 implies
\[
|S| + |I_0| \leq (2.5 + \epsilon)8C\sqrt{Kn_{\min} / \gamma_n} \sqrt{\sum_{k=1}^{K} n_k^2 \nu_k}.
\]

Combining this with condition (12) implies \( |S| + |I_0| < n_k \) for all \( k \) and hence \( G_k \cap (I_+ \backslash S) \neq \emptyset \). Therefore, for any two rows in \( G := I_+ \backslash S \), if they are in different clusters of \( \Theta \) then they must be in different clusters of \( \hat{\Theta} \) (otherwise, \( \|U'_{i \cdot} - U'_{j \cdot}\| \leq \|U''_{i \cdot} - \hat{\Theta}_{i \cdot} \hat{X}\| + \|\hat{\Theta}_{j \cdot} \hat{X} - U'_{j \cdot}\| < \sqrt{2} \)).

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 (22). The claimed result follows by choosing \( c = 8C \). \[\square\]
B Proof of Theorem 5.2

Throughout the proof we will use the following notation.

- For $i, j = 1, \ldots, n$, let $a_{ij}$ denote the $(i, j)$ entry of $A$ and $p_{ij} = E_{a_{ij}}$ the $(i, j)$ entry of $P$.
- Set $W = A - P$ and denote with $w_{ij}$ the $(i, j)$ entry of $W$.
- Set $p_{\text{max}} = \max_{ij} p_{ij}$ and $d = (np_{\text{max}}) \lor (c_0 \log n)$.
- For $i = 1, \ldots, n$, set $d_i$ to be the degree of node $i$ and $\bar{d}_i = d_i - E d_i$, the centered degree.
- For $t > 0$, let $S_t = \{v \in \mathbb{R}^n : \|v\|_2 \leq t\}$ be the Euclidean ball of radius $t$ and set $S = S_1$.

The proof of Theorem 5.2 is adapted from Feige & Ofek (2005), whose goal is to bound the second eigenvalue of a Erdős-Rényi random graph. The idea is to bound

$$\sup_{x \in S} |x^T (A - P) x|.$$  \hspace{1cm} (23)

The proof encompasses three major steps.

1. **Discretization.** We first reduce (23) 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 in $S$. 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, and the other part corresponds to the larger entries of $x$ or $y$, the heavy pairs. This idea is elaborated in Appendix B.1 and the validity of such a discretization is established in Lemma B.1.

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. This is done in Appendix B.2 and the main result of this step is Lemma B.2.

3. **Bounding the heavy pairs.** In the final step, the contribution from the heavy pairs 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. The details are presented in Appendix B.3. The result is summarized in Lemma B.5.

The claimed result of Theorem 5.2 follows immediately by combining Lemmas B.1, B.2 and B.5, which we state and prove below.

B.1 Discretization

Fix $\delta \in (0, 1)$, for example $\delta = 1/2$, and define

$$T = \left\{ x = (x_1, \ldots, x_n) \in S : \sqrt{n} x_i / \delta \in \mathbb{Z}, \forall i \right\},$$

where $\mathbb{Z}$ denotes the set of all integers. In other words, $T$ consists of all grid points of size $\delta / \sqrt{n}$ within the unit ball in $\mathbb{R}^n$. The next result, which is similar to Lemma 2.3 of Feige & Ofek (2005), quantifies the loss in accuracy when, in problem (23), the optimization is carried over $T$ instead of $S$.
Lemma B.1. \( S_{1-\delta} \subseteq \text{convhull}(T) \). As a consequence, for all \( W \in \mathbb{R}^{n \times n} \),
\[
\|W\|_2 \leq (1 - \delta)^{-2} \sup_{x,y \in T} |x^T Wy| .
\]

**Proof.** For any \( u \in S_{1-\delta} \) consider the cube of size \( \delta/\sqrt{n} \) that contains \( u \). The diameter of the cube is \( \delta \), so the entire cube is contained in \( S \). As a result, all vertices of this cube are in \( T \) and hence \( u \in \text{convhull}(T) \).

For any \( u \in S \), let \( (1 - \delta)u = \sum_i a_i x_i \) be the convex combination of points in \( T \). Then
\[
|u^T Wu| = (1 - \delta)^{-2}[(1 - \delta)u]^TW[(1 - \delta)u]
\leq (1 - \delta)^{-2} \sum_{ij} a_i a_j |x_i^T W x_j| \leq (1 - \delta)^{-2} \sup_{x,y \in T} |x^T Wy| .
\]

For any pair of vectors \( x, y \in T \), we have
\[
x^T (A - P)y = \sum_{1 \leq i,j \leq n} x_i y_j (a_{ij} - p_{ij}).
\]

We then split the pairs \( (x_i, y_j) \) into **light pairs**
\[
\mathcal{L} = \mathcal{L}(x,y) := \{(i,j) : |x_i y_j| \leq \sqrt{d/n}\},
\]
and into **heavy pairs**
\[
\bar{\mathcal{L}} = \bar{\mathcal{L}}(x,y) := \{(i,j) : |x_i y_j| > \sqrt{d/n}\}.
\]

We remark that the definitions of \( \mathcal{L} \) and \( \bar{\mathcal{L}} \) depend on \( x, y \), but we occasionally will omit this dependence in our notation for simplicity.

### B.2 Bounding the light pairs

The next Lemma bounds the contribution of the light pairs.

**Lemma B.2.**
\[
\mathbb{P} \left( \sup_{x,y \in T} \left| \sum_{(i,j) \in \mathcal{L}(x,y)} x_i y_j w_{ij} \right| \geq c\sqrt{d} \right) \leq \exp \left[ - \left( \frac{c^2}{4 + \frac{7}{3}} - 2 \log \left( \frac{7}{\delta} \right) \right) n \right]
\]

**Proof.** Let \( u_{ij} = x_i y_j \mathbf{1}(|x_i y_j| \leq \sqrt{d/n}) + x_j y_i \mathbf{1}(|x_j y_i| \leq \sqrt{d/n}) \), for \( i,j = 1, \ldots, n \). Then the contribution to \( x^T W y \) from the light pairs can be written as
\[
\sum_{1 \leq i,j \leq n} w_{ij} u_{ij}.
\]

Because \( |u_{ij}| \leq 2\sqrt{d}/n \), each term in the summand has mean zero and is bounded in absolute value by \( 2\sqrt{d}/n \). Apply Bernstein’s inequality
\[
\mathbb{P} \left[ \left| \sum_{i<j} w_{ij} u_{ij} \right| \geq c\sqrt{d} \right] \leq 2 \exp \left( - \frac{\frac{1}{2} c^2 d}{\sum_{i<j} p_{ij}(1 - p_{ij}) u_{ij}^2 + \frac{2\sqrt{d}}{n} c\sqrt{d}} \right)
\]

19
The other three cases are similar.

can be bounded by high probability. The claimed result follows from the union bound.

B.3 Bounding the heavy pairs

Next we will show that the contribution of heavy pairs,

$$\sup_{x,y \in T} \left| \sum_{(i,j) \in \mathcal{L}(x,y)} x_i y_j w_{ij} \right|,$$

can be bounded by \( cvd \) with high probability for some universal constant \( c \). Indeed, observe that

$$\left| \sum_{(i,j) \in \mathcal{L}} x_i y_j p_{ij} \right| = \sum_{(i,j) \in \mathcal{L}} \frac{x_i^2 y_j^2}{x_i y_j} p_{ij} \leq \frac{n}{\sqrt{d}} p_{\max} \sum_{(i,j) \in \mathcal{L}} x_i^2 y_j^2 \leq \sqrt{d} \sum_{i,j} x_i^2 y_j^2 \leq \sqrt{d},$$

where the second inequality comes from the definition of \( \mathcal{L} \). Therefore, it suffices to show that

$$\sum_{(i,j) \in \mathcal{L}} x_i y_j a_{ij} = O(\sqrt{d}),$$

with high probability.

We will focus on the heavy pairs \((i, j)\) such that \(x_i > 0, y_j > 0\) and, accordingly, set

$$\mathcal{L}_1 = \{(i, j) \in \mathcal{L} : x_i > 0, y_j > 0\}.$$

The other three cases are similar.

We will need the following notation.

- \( I_1 = \left\{ i : \frac{\delta}{\sqrt{n}} \leq x_i \leq \frac{2\delta}{\sqrt{n}} \right\}, \)
- \( I_s = \left\{ i : \frac{\delta}{\sqrt{n}} 2^{s-1} < x_i \leq \frac{\delta}{\sqrt{n}} 2^s \right\} \) for \( s = 2, 3, \ldots, \left[ \log_2 \sqrt{n} \right] \).
- \( J_1 = \left\{ j : \frac{\delta}{\sqrt{n}} \leq y_j \leq \frac{2\delta}{\sqrt{n}} \right\}, \)
- \( J_t = \left\{ j : \frac{\delta}{\sqrt{n}} 2^{t-1} < y_j \leq \frac{\delta}{\sqrt{n}} 2^t \right\} \) for \( t = 2, 3, \ldots, \left[ \log_2 \sqrt{n} \right] \).
- \( e(I, J) \): the number of distinct edges between node sets \( I \) and \( J \). If \( I \) and \( J \) are disjoint, then \( e(I, J) = \sum_{i \in I, j \in J} a_{ij} \). If \( I \cap J \neq \emptyset \) then \( e(I, J) = \sum_{(i,j) \in (I \times J) \setminus (I \cap J^2)} a_{ij} + \sum_{(i,j) \in (I \cap J^2), i < j} a_{ij} \).
\[ \mu(I, J) = \mathbb{E}e(I, J), \quad \tilde{\mu}(I, J) = p_{\text{max}}|I||J|. \]
We will use \( \mu \) and \( \tilde{\mu} \) for convenience when there is no need to specify the dependence on \( I \) and \( J \).

- \( \lambda_{st} = e(I_s, J_t)/\tilde{\mu}_{st} \), where \( \tilde{\mu}_{st} = \tilde{\mu}(I_s, J_t) \).
- \( \alpha_s = |I_s|2^{2s}/n, \beta_t = |J_t|2^{2t}/n, \sigma_{st} = \lambda_{st}\sqrt{d2^{-(s+t)}}. \)

We begin by defining two events that will play a key role in the proof and then showing that their probability approaches 1 as \( n \) grows.

**Lemma B.3** (Bounded degree). For \( c > 0 \), there exists constant \( c_1 = c_1(c) \) such that with probability at least \( 1 - n^{-c} \), \( d_i \leq c_1d \) for all \( i \).

**Lemma B.4** (Bounded Discrepancy). For \( c > 0 \), there exists constants \( c_2 = c_2(c) \) and \( c_3 = c_3(c) \), both larger than 1, such that with probability at least \( 1 - 2n^{-c} \), for any \( I, J \subseteq [n] \) with \( |I| \leq |J| \), at least one of the following holds:

1. \( \frac{e(I, J)}{\mu(I, J)} \leq cc_2 \);

2. \( e(I, J) \log \frac{e(I, J)}{\mu(I, J)} \leq c_3|J| \log |J| \).

Lemmas B.3 and B.4 are proved in Appendix B.4. Below we will assume that the events given in Lemmas B.3 and B.4 both hold. That is, we focus on realizations of \( A \) for which both the degree and discrepancy are bounded, in the sense of Lemmas B.3 and B.4.

Simple algebra gives

\[
\sum_{(i,j) \in \mathcal{L}_1} x_iy_ja_{ij} \leq 2^s \delta 2^{t+\delta} \sum_{s+t : s+2t \geq \sqrt{d}} e(I_s, J_t) \frac{2^s \delta 2^{t+\delta}}{\sqrt{n} \sqrt{n}}
\]

\[
= 2\delta^2 \sqrt{d} \sum_{s+t : s+2t \geq \sqrt{d}} \alpha_s \beta_t \sigma_{st} \tag{24}
\]

where the factor of 2 comes from the fact that each distinct edge between \( I_s \) and \( J_t \) appears at most twice in the summation.

We bound this sum by splitting the pairs of \((s, t)\) into twelve categories. Let \( \mathcal{C} = \{(s, t) : 2^{s+t} \geq \sqrt{d}, |I_s| \leq |J_t|\} \) and define the following sets:

- \( \mathcal{C}_1 = \{(s, t) \in \mathcal{C} : \sigma_{st} \leq 1\} \).
- \( \mathcal{C}_2 = \{(s, t) \in \mathcal{C}\setminus \mathcal{C}_1 : \lambda_{st} \leq cc_2\} \).
- \( \mathcal{C}_3 = \{(s, t) \in \mathcal{C}\setminus (\mathcal{C}_1 \cup \mathcal{C}_2) : 2^s \geq \sqrt{d}2^t\} \).
- \( \mathcal{C}_4 = \{(s, t) \in \mathcal{C}\setminus (\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3) : \log \lambda_{st} > \frac{1}{2}[2t \log 2 + \log(1/\beta_t)]\} \).
- \( \mathcal{C}_5 = \{(s, t) \in \mathcal{C}\setminus (\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_4) : 2t \log 2 \geq \log(1/\beta_t)\} \).
- \( \mathcal{C}_6 = \{(s, t) \in \mathcal{C}\setminus (\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_4 \cup \mathcal{C}_5)\} \).

The other six categories can be defined using a similar partition of \( \mathcal{C}' = \{(s, t) : 2^{s+t} \geq \sqrt{d}, |I_s| > |J_t|\} \)
by switching the roles of \((I, s, \alpha)\) and \((J, t, \beta)\).

We now analyze separately each of the six cases. Towards that end, we will repeatedly make use of the following simple facts

\[
\sum_s \alpha_s \leq \sum_i |2x_i/\delta|^2 \leq 4\delta^{-2} \quad \text{and} \quad \sum_t \beta_t \leq 4\delta^{-2}.
\]

**Pairs in** \(C_1\). In this case we get the bound

\[
\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1((s, t) \in C_1) \leq \sum_{(s,t)} \alpha_s \beta_t 1((s, t) \in C_1)
\leq \sum_{s,t} \alpha_s \beta_t = 4\|x\|^2_2|y\|^2_2 \delta^{-2} \leq 16\delta^{-4}.
\]

**Pairs in** \(C_2\). This is the first case claimed in the discrepancy lemma. In this case

\[
\sigma_{st} = \lambda_{st} \sqrt{d} 2^{-(s+t)} \leq \lambda_{st} \leq e c_2.
\]

Therefore,

\[
\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1((s, t) \in C_2) \leq \sum_{(s,t)} \alpha_s \beta_t e c_2 1((s, t) \in C_2)
\leq e c_2 \sum_{s,t} \alpha_s \beta_t = e c_2 4\|x\|^2_2|y\|^2_2 \delta^{-2} \leq e c_2 16\delta^{-4}.
\]

**Pairs in** \(C_3\). In this case \(2^{s-t} \geq \sqrt{d}\). Also by the bounded degree Lemma B.3, we have

\[
e(I_s, J_t) \leq c_1 |I_s|d,
\]

and hence \(\lambda_{st} \leq c_1 n/|J_t|\). Thus,

\[
\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1((s, t) \in C_3) = \sum_s \alpha_s \sum_t \beta_t \sigma_{st} 1((s, t) \in C_3)
= \sum_s \alpha_s \sum_t |J_t| 2^{2t} n \lambda_{st} \sqrt{d} 2^{-(s+t)} 1((s, t) \in C_3)
\leq \sum_s \alpha_s \sum_t |J_t| 2^{2t} c_1 n |J_t| \sqrt{d} 2^{-(s+t)} 1((s, t) \in C_3)
\leq c_1 \sum_s \alpha_s \sum_t \frac{\sqrt{d}}{2^{-(s-t)}} 1((s, t) \in C_3) \leq 2c_1 \sum_s \alpha_s \leq 8c_1 \delta^{-2},
\]

where the first inequality uses \(\lambda_{st} \leq c_1 n/|J_t|\), the second inequality follows from the definition of \(C_3\), and the third inequality follows from the fact that the non-zero summands over \(t\) are all bounded by 1 and is a geometric sequence.

In order to bound the pairs in \(C_4\), \(C_5\), and \(C_6\), we will rely on the second case described in the bounded discrepancy lemma, which we can rewrite in an equivalent form as

\[
\lambda_{st}|I_s||J_t|d n \log \lambda_{st} \leq c_3 |J_t| \log \frac{2^{2t}}{|J_t|}.
\]
Re-arranging, this is in turn equivalent to
\[
\sigma_s \alpha_s \log \lambda_{st} \leq c_3 \frac{2^{s-t}}{\sqrt{d}} \left[2t \log 2 + \log \beta_t^{-1}\right],
\]
which is particularly convenient for our purposes.

**Pairs in \( C_4 \).** The inequality \( \log \lambda_{st} > \frac{1}{t}[2t \log 2 + \log(1/\beta_t)] \) and (25) imply that \( \alpha_s \sigma_{st} \leq 4c_3 2^{s-t}/\sqrt{d} \). Then
\[
\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1((s,t) \in C_4) = \sum_t \beta_t \sum_s \alpha_s \sigma_{st} 1((s,t) \in C_4) \\
\leq 4c_3 \sum_t \beta_t \sum_s 2^{s-t}/\sqrt{d} 1((s,t) \in C_4) \leq 8c_3 \sum_t \beta_t \leq 32c_3\delta^{-2},
\]
where the first inequality uses the property of \( C_4 \) which implies that \( \alpha_s \sigma_{st} \leq 4c_3 2^{s-t}/\sqrt{d} \), and the second inequality follows from the fact that the nonzero summand over \( s \) is a geometric sequence bounded by 1 because \( (s,t) \notin C_3 \).

**Pairs in \( C_5 \).** In this case we have \( 2t \log 2 \geq \log \beta_t^{-1} \). Also because \( (s,t) \notin C_4 \), we have \( \log \lambda_{st} \leq 4^{-1}[2t \log 2 + \log \beta_t^{-1}] \leq t \log 2 \). Thus \( \lambda_{st} \leq 2^t \). On the other hand, because \( (s,t) \notin C_1 \), \( 1 \leq \sigma_{st} = \lambda_{st} \sqrt{d} 2^{-(s+t)} \leq \sqrt{d} 2^{-s} \). Thus \( 2^s \leq \sqrt{d} \).

Because \( (s,t) \notin C_2 \), we have \( \log \lambda_{st} \geq 1 \), combining with \( 2t \log 2 \geq \log \beta_t^{-1} \), eq. (25) implies that
\[
\sigma_{st} \alpha_s \leq c_3 \frac{2^{s-t}}{\sqrt{d}} 4t \log 2.
\]
Then,
\[
\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1((s,t) \in C_5) = \sum_t \beta_t \sum_s \alpha_s \sigma_{st} 1((s,t) \in C_5) \\
\leq \sum_t \beta_t \sum_s c_3 \frac{2^{s-t}}{\sqrt{d}} 4t \log 2 1((s,t) \in C_5) \\
\leq 4c_3 \log 2 \sum_t \beta_t 2^{t} \sum_s c_3 \frac{2^{s}}{\sqrt{d}} 1((s,t) \in C_5) \\
\leq 2c_3 \log 2 \sum_t \beta_t \leq 8c_3\delta^{-2}.
\]

**Pairs in \( C_6 \).** We have \( 2t \log 2 < \log \beta_t^{-1} \). Because \( (s,t) \notin C_4 \), we have \( \log \lambda_{st} \leq \frac{1}{t} \log \beta_t^{-1} \leq \log \beta_t^{-1} \) where the last inequality comes from the fact that \( \log \lambda_{st} \geq 1 \) because \( (s,t) \notin C_2 \).
\[
\sum_{(s,t)} \alpha_s \beta_t \sigma_{st} 1((s,t) \in C_6) = \sum_s \alpha_s \sum_t \beta_t \lambda_{st} \sqrt{d} 2^{-(s+t)} 1((s,t) \in C_6) \\
\leq \sum_s \alpha_s \sum_t \sqrt{d} 2^{-(s+t)} 1((s,t) \in C_6) \\
\leq 2 \sum_s \alpha_s \leq 8\delta^{-2}.
\]

Now we have proved the following lemma.
Lemma B.5 (Heavy pair bound). For any given $c > 0$, there exists a constant $C$ depending only on $c$ such that

$$\sup_{x,y \in T} \left| \sum_{(i,j) \in L} x_i y_j w_{ij} \right| \leq C \sqrt{d},$$

with probability at least $1 - 2n^{-c}$.

B.4 Proof of auxiliary results

Now we prove Lemmas B.3 and B.4.

Proof of Lemma B.3. For a fixed node $i$, using Bernstein’s inequality

$$\Pr(d_i \geq c_1 d) \leq \Pr \left( \sum_{j=1}^{n} w_{ij} \geq (c_1 - 1)d \right) \leq \exp \left[-\frac{\frac{1}{2} (c_1 - 1)^2 d^2}{\sum_{j=1}^{n} p_{ij} (1 - p_{ij}) + \frac{1}{3} (c_1 - 1)d}\right] \leq n^{-\frac{3c_0(c_1 - 1)^2}{2c_1 + 4}},$$

where the last inequality uses the assumption $d \geq c_0 \log n$.

Proof of Lemma B.4. We will assume throughout that the event of bounded degree with constant $c_1$ described in Lemma B.3 holds.

If $|J| \geq n/e$, then the bounded degree implies that $\frac{e(I,J)}{d(I,J)/n} \leq \frac{|J|c_1 d}{d(I,J)/e} \leq c_1 e$.

On the other hand, if $|J| < n/e$, let $k > 1$ be a positive number to be chosen later. Denote $s(I,J)$ the set of all possible distinct edges $(i,j)$ between $I$ and $J$. Using Corollary A.1.10 of Alon & Spencer (2004)

$$\Pr [e(I,J) \geq k\bar{\mu}(I,J)] \leq \Pr \left[ \sum_{(i,j) \in s(I,J)} (a_{ij} - p_{ij}) \geq k\bar{\mu}(I,J) - \sum_{(i,j) \in s(I,J)} p_{ij} \right] \leq \Pr \left[ \sum_{(i,j) \in s(I,J)} w_{ij} \geq (k - 1)\bar{\mu}(I,J) \right] \leq \exp \left[ (k - 1)\bar{\mu} - k\bar{\mu} \log k \right] \leq \exp \left[ -\frac{1}{2} (k \log k) \bar{\mu} \right],$$

where the last inequality holds for $k \geq 8$.

For a given number $c_3 > 0$, define $t(I,J)$ as the unique value of $t$ such that $t \log t = \frac{c_3 |J|}{\bar{\mu}(I,J)} \log \frac{n}{|J|}$. Let $k(I,J) = \max\{8, t(I,J)\}$.

Then,

$$\Pr [e(I,J) \geq k(I,J)\bar{\mu}(I,J)] \leq \exp \left[ -\frac{1}{2} \bar{\mu}(I,J) k(I,J) \log k(I,J) \right] \leq \exp \left[ -\frac{1}{2} c_3 |J| \log \frac{n}{|J|} \right].$$
Therefore,

\[
\Pr \left[ \exists (I, J) : |I| \leq |J| \leq \frac{n}{e}, e(I, J) \geq k(I, J)\bar{\mu}(I, J) \right] \\
\leq \sum_{I, J: |I| \leq |J| \leq n/e} \exp \left[ -\frac{1}{2} c_3 |J| \log \frac{n}{|J|} \right] \\
\leq \sum_{h, g: 1 \leq h \leq g \leq n/e} \sum_{I, J: |I| = h, |J| = g} \exp \left[ -\frac{1}{2} c_3 g \log \frac{n}{g} \right] \\
= \sum_{h, g: 1 \leq h \leq g \leq n/e} \binom{n}{h} \binom{n}{g} \exp \left[ -\frac{1}{2} c_3 g \log \frac{n}{g} \right] \\
\leq \sum_{h, g: 1 \leq h \leq g \leq n/e} \left( \frac{ne}{h} \right)^h \left( \frac{ne}{g} \right)^g \exp \left[ -\frac{1}{2} c_3 g \log \frac{n}{g} \right] \\
= \sum_{h, g: 1 \leq h \leq g \leq n/e} \exp \left[ -\frac{1}{2} c_3 g \log \frac{n}{g} + h \log \frac{n}{h} + g \log \frac{n}{g} + g \right] \\
\leq \sum_{h, g: 1 \leq h \leq g \leq n/e} \exp \left[ -\frac{1}{2} c_3 g \log \frac{n}{g} + 2g \log \frac{n}{g} + 2g \right] \\
\leq \sum_{h, g: 1 \leq h \leq g \leq n/e} \exp \left[ -\frac{1}{2} (c_3 - 8) g \log \frac{n}{g} \right] \\
\leq \sum_{h, g: 1 \leq h \leq g \leq n/e} n^{-\frac{1}{2}(c_3-8)} \\
\leq n^{-\frac{1}{2}(c_3-12)},
\]

where the inequalities repeatedly use the assumption that \( h \leq g \leq n/e \) and the fact that \( t \log \frac{t}{t} \) is increasing on \([1, n/e]\).

As a result, with probability at least \( 1 - n^{-\frac{1}{2}(c_3-12)} \), we have \( c(I, J) \leq k(I, J)\bar{\mu}(I, J) \), for all \( |I| \leq |J| \leq n/e \). As a final step, we further divide the set of pairs \((I, J)\) satisfying \( |I| \leq |J| \leq n/e \) into two groups by the value of \( k(I, J) \). For the pairs for which \( k(I, J) = 8 \), we get

\[
e(I, J) \leq k(I, J)\bar{\mu}(I, J) = 8\bar{\mu}(I, J).
\]

For the all the others pairs \( k(I, J) = t(I, J) > 8 \), and we have \( \frac{e(I, J)}{\bar{\mu}(I, J)} \leq t(I, J) \). Thus

\[
\frac{e(I, J)}{\bar{\mu}(I, J)} \log \frac{e(I, J)}{\bar{\mu}(I, J)} \leq t(I, J) \log t(I, J) = \frac{c_3 |J|}{\bar{\mu}(I, J)} \log \frac{n}{|J|},
\]

which implies that

\[
e(I, J) \log \frac{e(I, J)}{\bar{\mu}(I, J)} \leq c_3 |J| \log \frac{n}{|J|}.
\]

The desired claim follows by letting \( c_2 = \max\{c_1, 8\} \), and \( c_3 = 2c + 12 \).

\[ \square \]

**References**

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). Np-hardness of euclidean sum-of-squares clustering. *Machine Learning, 75*(2), 245–248.

25
Alon, N., & Spencer, J. H. (2004). *The probabilistic method*. Wiley. com.

Amini, A. A., Chen, A., Bickel, P. J., & Levina, E. (2012). Pseudo-likelihood methods for community detection in large sparse networks. *arXiv preprint arXiv:1207.2340*.

Anandkumar, A., Ge, R., Hsu, D., & Kakade, S. M. (2013). A tensor spectral approach to learning mixed membership community models. *arXiv preprint arXiv:1302.2684*.

Awasthi, P., & Sheffet, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, (pp. 37–49). Springer.

Balakrishnan, S., Xu, M., Krishnamurthy, A., & Singh, A. (2011). Noise thresholds for spectral clustering. In *In Advances in Neural Information Processing Systems 25*.

Bhatia, R. (1997). *Matrix analysis*, vol. 169. Springer.

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

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

Channarond, A., Daudin, J.-J., & Robin, S. (2012). Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electronic Journal of Statistics, 6*, 2574–2601.

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

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

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

Choi, D. S., Wolfe, P. J., & Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika, 99*(2), 273–284.

Chung, F., & Radcliffe, M. (2011). On the spectra of general random graphs. *the electronic journal of combinatorics, 18*(P215), 1.

Coja-Oghlan, A. (2010). Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing, 19*, 227–284.

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

Deshpande, Y., & Montanari, A. (2013). Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *arXiv preprint arXiv:1304.7047*. 26
Feige, U., & Ofek, E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms, 27*(2), 251–275.

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

Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning, 2*(2), 129–233.

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

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

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

Kolaczyk, E. D. (2009). *Statistical analysis of network data*. Springer.

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

Kumar, A., & Kannan, R. (2010). Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, (pp. 299–308). IEEE.

Kumar, A., Sabharwal, Y., & Sen, S. (2004). A simple linear time \((1 + \epsilon)\)-approximation algorithm for k-means clustering in any dimensions. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, (pp. 454–462). IEEE.

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

Lu, L., & Peng, X. (2012). Spectra of edge-independent random graphs. arXiv preprint arXiv:1204.6207.

Lyzinski, V., Sussman, D., Tang, M., Athreya, A., & Priebe, C. (2013). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. arXiv preprint arXiv:1310.0532.

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

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

Mossel, E., Neeman, J., & Sly, A. (2012). Stochastic block models and reconstruction. arXiv preprint arXiv:1202.1499.

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

Newman, M. (2009). *Networks: an introduction*. Oxford University Press.
Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E, 69*(2), 026113.

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems, 2*, 849–856.

Qin, T., & Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. *arXiv preprint arXiv:1309.4111*.

Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics, 39*, 1878–1915.

Sarkar, P., & Bickel, P. (2013). Role of normalization in spectral clustering for stochastic blockmodels. *arXiv preprint arXiv:1310.1495*.

Sussman, D. L., Tang, M., Fishkind, D. E., & Priebe, C. (2013). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *arXiv preprint arxiv:1310.0532*.

Sussman, D. L., Tang, M., Fishkind, D. E., & Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association, 107*(499), 1119–1128.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics, 12*(4), 389–434.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing, 17*(4), 395–416.

Vu, V. Q., & Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Annals of Statistics, 41*(6), 2905–2947.

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