On Spectral Clustering for Sparse Stochastic Block Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We analyze the performance of a practical spectral clustering algorithm for community extraction in the stochastic block model. The procedure performs k-means clustering on the leading eigenvectors of the adjacency matrix of the observed network. We provide sufficient conditions for consistent community recovery in terms of the overall sparsity, the degree of separation among communities, and the imbalance among the sizes the communities. We show that the algorithm can recover the hidden communities with vanishing misclustering rate even when the expected node degrees grow only logarithmically in the size of the network. We demonstrate rates that are comparable or better than those reported in most existing work, which are often based on more computationally demanding algorithms.

1 Introduction

In recent years, modern technology has enabled many new forms of measurement and collection of complex data. An important subcategory is network or relational data. In a network data set, the recorded values are not attributes of individuals from the same population, but the interactions between pairs of individuals in the population. 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 other fields. A network data set can be represented by the network *adjacency matrix* $A = (a_{ij})_{1\le i,j\le n}$, a symmetric binary matrix with zero diagonal values with each entry indicating if there is a connection between node i and node j. Here $\mathcal{V} = \{1, 2, ..., n\}$ is the set of all nodes in the sample. A review of modeling and inference on network data can be found in recent books [14, 10].

An important inference task for network data is the identification of communities, where, loosely speaking, a community is a subset of nodes in the network that have an higher average degree of connectivity within themselves compared to the remaining nodes. There are different methods for finding communities from a given adjacency matrix. In statistics and machine learning, a simple and nice mixture model is proposed by [11], with the name “stochastic block model” (SBM). In a SBM, the nodes are partitioned into K disjoint communities, where the chance of there being an edge between node i and node j is determined only by the community membership of i and j. Moreover, it is assumed that the edges are independently generated given the membership.

The SBM has been the focus of much research effort on network community detection because it has a simple and intuitive mathematical structure that allows for rigorous analysis. Some extensions of SBM, such as the degree corrected block model [13] and the mixed membership model [1] can be used to approximate a wide range of real network data. Such a model based approach to network community extraction also opens the possibility of statistical inferences such as goodness-of-fit test and confidence intervals. In the statistics literature, the problem of consistent community estimation under the stochastic block model and variants thereof is studied in [4, 20, 5]. In computer science and machine learning, the network community detection problem is closely related to the graph partitioning and planted partition model (a special case of SBM), where the spectral clustering method is very popular [17, 19].

We study the community recovery problem in the stochastic block model using a simple form of spectral clustering for network data. The applicability of spectral methods for community extraction was initially shown by [20], who demonstrated that spectral clustering using normalized graph Laplacian gives consistent community detection for dense graphs where the node degrees are of order $n/\log n$. Our result gives an affirmative answer to an open question raised by [20], showing that spectral clustering methods can consistently detect the communities for very sparse block models. We investigate the misclustering rate in terms of the overall sparsity, the separation of connectivity between blocks, and the maximum and minimum block sizes. In particular, we show that a simpler spectral clustering method can give consistent community detection even when the maximum node degree is of order $\log n$, which is a weaker requirement than most existing results. Such an improvement may be due to a new argument used in this paper, together with the fact that we focus on approximate community recovery (vanishing Hamming error rate), rather than the more stringent exact recovery considered in many recent works [4, 7, 6]. The dependence on the community separation in our result is comparable to other state-of-art results such as [7, 6]. Our result explicitly keeps track of all logarithm terms, because, as mentioned in [4], $\log n$ seems to be a critical rate of the average degree in the sparse regime of stochastic block models.

The method and some of the analysis of this paper are inspired by [12], who studies community extraction in the (dense) degree corrected stochastic block model ([13]). The method proposed there simply performs a k-means clustering algorithm on the leading eigenvectors of the adjacency matrix, without using any graph Laplacian. This procedure does not require any tuning parameters once the number of communities, K, is known. Unlike some existing works (including [12, 6]) where the matrix Bernstein's inequality is the main technical tool, our analysis relies additionally on a new variational characterization of the principal components recently developed by [23], together with a sharp spectral bound of random matrices with block structured expectation ([9]). These arguments give a better dependence on the sparsity and community separation when K is small and require a much weaker eigengap condition than in [20] and [12].

Contribution of this paper The main contributions of this paper are the following.

1. We analyze perhaps the simplest form of spectral clustering for community detection in stochastic block models with novel arguments under general conditions which allow the model to be very sparse.
2. Our misclustering rate in Hamming distance explicitly takes into account the overall sparsity, the degree of separation among communities and community size imbalance.
3. We show that consistent community detection is achievable when the maximum degree is of order $\log n$, while the dependence on block separation is better than other state-of-art results when the number of blocks is small.

Related work Consistency of spectral clustering under stochastic block models has so far been mostly focusing on very dense graphs. The first work in this direction is [20], where the normalized graph Laplacian is considered and the network is assumed to be dense. Extensions to (dense) degree corrected block models is reported in [12] under Hamming distance error rate. [6] studies community detection for extended planted partition (a special case of degree corrected stochastic block model) using the random walk graph Laplacian.

Other approaches to community detection for stochastic block models include optimization over block partitions such as modularity methods [18] and likelihood methods [4, 8]; matrix optimization [7]; and tensor spectral methods [2]. The performance is commonly assessed by the minimum average degree and the separation between intra- and inter-block connectivity required for consistent community detection. The dependence on average degree is complicated due to the interaction of all the factors in the model. The likelihood modularity method proposed in [4] can succeed when the average degree grows faster than $\log n$. The convex optimization method in [7] allows the average degree to be as small as $\log^4 n$. In the graph partitioning literature, [9] obtained a very sharp result for a spectral method (other than spectral clustering) that essentially requires the average degree to be larger than K^4 where K is the number of blocks. The algorithm and analysis presented in this paper are much simpler and more implementable than those in [9].

2 Stochastic block model and spectral clustering

In the stochastic block model [11], the adjacency matrix $A = (a_{ij})_{1 \le i,j \le n}$ consists of independent upper diagonal entries $a_{ij} \sim \mathrm{Bernoulli}(p_{ij})$ for $1 \le i < j \le n$, with $a_{ji} = a_{ij}$ and $a_{ii} = 0$ for all i. Additionally, the stochastic block model assumes that the edge probabilities $(p_{ij} : 1 \le i < j \le n)$ come from the entries of a matrix $P = (p_{ij})_{1 \le i,j \le n}$, which has a *block structure*: $P = \Psi B \Psi^T$, where $B = (b_{k\ell})_{1 \le k,\ell \le K}$ is a $K \times K$ symmetric matrix with $b_{k\ell} \in [0, 1]$ for all $1 \le k, \ell \le K$, and Ψ is an $n \times K$ membership matrix, with each row having one entry being 1 and others being 0. It is usually assumed that K is much smaller than n.

A stochastic block model is specified by the pair (Ψ, B). It naturally partitions the set of nodes into K communities through the membership matrix Ψ. Formally, for each $i \in [n]$, let $g(i) \in [K]$ be such that $\Psi_{i,g(i)} = 1$. Then $g(i)$ indicates the community (or block) that node i belongs to. The model further assumes that the probability of an edge between nodes i and j is $b_{g(i)g(j)}$. Therefore, the model uses B to capture the average degree of connectivity among and within the communities, and the interaction between two nodes is determined by their community membership and the corresponding entry in B. The community identification problem is to recover the membership matrix Ψ (up to column permutations) from a realization of A. If the communities are correctly recovered, estimating B is straightforward and can be done at a parametric rate.

Additional notation In the rest of this paper, $\|v\|_2$ denotes the Euclidean norm of a vector v. For any matrix $X \in \mathbb{R}^{n \times m}$, $\|X\|_F \equiv (\sum_{i,j} X_{ij}^2)^{1/2}$ denotes its Frobenius norm, $\|X\|_1 \equiv \sum_{i,j} |X_{ij}|$ its entry-wise ℓ_1 norm, and $\|X\|$ the operator norm. For any $G \subseteq [n]$, $X_{G\cdot}$ denotes the submatrix of X with row indices in G. In particular, for any $i \in [n]$, $X_{i\cdot}$ denotes the ith row of X. We use $\mathbb{M}^{n \times K}$ to denote the set of all $n \times K$ membership matrices (each row has one "1" and $K - 1$ "0"). Finally, $a_n = o(b_n)$ means that $\lim_{n\to\infty} |a_n/b_n| = 0$; $a_n = O(b_n)$ means $\limsup_{n\to\infty} |a_n/b_n| < \infty$; $a_n = \omega(b_n)$ means $b_n = o(a_n)$; and $a_n = \Omega(b_n)$ means $b_n = O(a_n)$.

A spectral clustering algorithm We consider the simple algorithm proposed in [12], whose intuition is straightforward. The adjacency matrix A can be viewed as a noisy version of the underlying matrix P with independent zero-mean additive noises. Let ψ_k be the kth column of Ψ, and $\Delta = \mathrm{diag}(\|\psi_1\|_2, ..., \|\psi_K\|_2)$. Then it is easy to verify that the columns of $\tilde\Psi \equiv \Psi\Delta^{-1}$ are orthonormal. Now let $\tilde B = \Delta B \Delta$ with eigen-decomposition $\tilde B = QDQ^T$, where $D = \mathrm{diag}(d_1, ..., d_K)$ satisfies $|d_1| \ge |d_2| \ge ... \ge |d_K| \ge 0$, and $Q \in \mathbb{R}^{K\times K}$ is an orthonormal matrix. Then we can write

$$P = \Psi B \Psi^T = \tilde\Psi \tilde B \tilde\Psi^T = (\tilde\Psi Q) D (\tilde\Psi Q)^T, \tag{1}$$

which is the eigen-decomposition of P because D is diagonal and $(\tilde\Psi Q)(\tilde\Psi Q)^T = \tilde\Psi\tilde\Psi^T = I_K$. Note that $\tilde\Psi$ has only K distinct rows. As a result $\tilde\Psi Q$ also has only K distinct rows. Specifically, these distinct rows are $\{Q_{k\cdot}/\|\psi_k\|_2, k \in [K]\}$.

Below (see Lemma 5 in Section 3) we show that the eigenvectors of A are close to those of P, then one can recover the community membership by grouping the rows of the leading eigenvectors of A, which is precisely spectral clustering. The details of the algorithm are given below.

Algorithm 1: Simple Spectral Clustering

Input: Adjacency matrix A, number of blocks K, (optional) approximation parameter ϵ for k-means subroutine
Output: A membership matrix $\hat\Psi$.

1. Let $\hat U \hat D \hat U^T$ be the leading K-dimensional eigen-decomposition of A.
2. Output the clustering given by any $(1+\epsilon)$-approximate k-means algorithm on rows of $\hat U$.

Remark 1. *To give a precise formulation of the k-means problem, consider the following optimization problem,*

$$\min_{\Psi \in \mathbb{M}^{n\times K}, X \in \mathbb{R}^{K\times K}} \|\Psi X - \hat{U}\|_F^2 \,. \tag{2}$$

It is well-known that finding the exact solution to (2) *is NP-hard. But efficient algorithms have been developed to find constant factor approximations (see, e.g., [15]). We will investigate the performance of spectral clustering for both exact and approximate solutions to* (2).

Algorithm 1 is attractive because of its simplicity. Obviously, its performance depends on the matrices B and Δ. [12] considered the case $\epsilon = 0$ and showed that if $b_{k\ell} = \Omega(1)$ for all $1 \le k, \ell \le K$ then the misclustering rate of this algorithm goes to zero with high probability, provided that the smallest cluster size is large enough and the eigengap of B is bounded away from zero. In the following section, we give a better quantification of the misclustering rate in terms of $\max_{k,\ell} b_{k\ell}$, $\min_k \|\psi_k\|_2$, K, and the smallest eigenvalue of B, which makes the method consistent even when $b_{k\ell} = \Omega(\log n/n)$.

3 Analysis

First, we introduce some notation and define some important quantities. Recall that the maximum expected degree in the network is bounded by $n \max_{1\le k,\ell\le K} b_{k\ell}$. In order to capture the overall sparsity of the block model, we consider the maximum blockwise connectivity

$$\alpha_n = \max_{1\le k,\ell \le K} b_{k\ell}\,,$$

which is allowed to change with n. By definition of α_n, the maximum entry of the scaled matrix $B_0 = B/\alpha_n$ is 1. Another important factor that determines the hardness of the community detection problem is the separation between different blocks. That is, the pairwise difference between the rows of B. Intuitively, if $B_{k\cdot}$ and $B_{\ell\cdot}$ are close, then it is hard to distinguish these two blocks. Here we use $\lambda_{\min}(B_0)$, the smallest absolute eigenvalue of B_0, as an indirect measure the block separation. To sum up, we use the following notation

$$B = \alpha_n B_0, \quad \max_{k,\ell} B_0(k,\ell) = 1, \quad \lambda_{\min}(B_0) = \lambda_n. \tag{3}$$

The quantities and scaling given in (3) separates out and emphasizes two fundamental aspects of the community detection problem: the overall sparsity (α_n), and the scaled block connectivity separation (λ_n). When α_n and λ_n are large, it is easier to recover the communities from noisy observations.

For $1 \le k \le K$, let n_k be the number of nodes in the kth block, and define

$$n_{\min} = \min_k n_k, \quad n_{\max} = \max_k n_k\,. \tag{4}$$

In the rest of this paper, we will characterize the performance of the spectral clustering method described above as a function of the parameters $(\alpha_n, \lambda_n, K, n_{\max}, n_{\min})$, which we allow to change with n.

3.1 Consistency of clustering

Recall that we seek to obtain a membership matrix $\hat{\Psi} \in \mathbb{M}^{n\times K}$ such that the total number of misclustered nodes is small. To this end, we define two types of consistency.

Definition 2. *We say* $\hat{\Psi}$ *is* overall consistent *if*

$$\mathrm{Err}_n(\hat{\Psi}, \Psi) \equiv (2n)^{-1} \min_{J\in\mathcal{P}_K} \|\hat{\Psi} - \Psi J\|_1 \to 0, \ \ as \ \ n \to \infty\,,$$

where $\mathcal{P}_K$ denotes the set of all $K \times K$ permutation matrices. Similarly, we say $\hat{\Psi}$ *is* blockwise consistent *if*

$$\lim_{n\to\infty} \min_{J\in\mathcal{P}_K} \max_{1\le k\le K} (2n_k)^{-1} \|\hat{\Psi}_{G_k\cdot} - \Psi_{G_k\cdot} J\|_1 = 0\,.$$

The notion of overall consistency is clearly weaker than that of blockwise consistency. Because both $\hat{\Psi}$ and Ψ are membership matrices, the matrix ℓ_1 norm used in the definition yields the (normalized) Hamming distance between the true and estimated community assignments. In both definitions, an error rate of 0 means perfect recovery, while an error rate of 1 means completely incorrect recovery.

Now we have our main results on the error bound of the spectral clustering algorithm. All the proofs are given in the Supplementary Material.

Theorem 3 (Error bounds of spectral clustering)**.** *Given any positive constants (a, r, ϵ), let A be a realization of stochastic block model (Ψ, B) such that*

$$\alpha_n \geq a\frac{\log n}{n}, \quad \frac{K(K \wedge \sqrt{\log n})^2 n}{\alpha_n \lambda_n^2 n_{\min}^2} < c_0^{-1}(a, r, \epsilon), \tag{5}$$

where c_0 is a deterministic function of (a, r, ϵ). Then with probability at least $1 - 2n^{-r}$, the output $\hat{\Psi}$ of Algorithm 1 satisfies, for some $K \times K$ permutation matrix J,

$$\sum_{k=1}^{K} \frac{\|\hat{\Psi}_{G_k\cdot} - \Psi_{G_k\cdot} J\|_1}{n_k} \leq \frac{2c_0(a, r, \epsilon)K(K \wedge \sqrt{\log n})^2 n}{\alpha_n \lambda_n^2 n_{\min}^2},$$

and as a consequence

$$\mathrm{Err}_n(\hat{\Psi}, \Psi) \leq \frac{c_0(a, r, \epsilon)K(K \wedge \sqrt{\log n})^2 n_{\max}}{\alpha_n \lambda_n^2 n_{\min}^2}.$$

If $\alpha_n \geq a(\log n)^4/n$, the above results hold when the term $K \wedge \sqrt{\log n}$ is replaced by 1.

The first condition in (5) is required to yield non-trivial bounds on $\|A - P\|$ (see Lemma 6), while the second condition is needed to guarantee a low misclustering error for the k-means procedure (see Lemma 7). In Section 3.2 we provide a detailed discussion and comparisons with the conditions assumed by other competing methods.

Remark 4. *The function c_0 in Theorem 3 is given in closed form in the Supplementary Material.*

The proof of Theorem 3 follows and refines the ideas outlined in [20, 12]. It consists of three major steps. The first step is to bound the distance from $\hat{U}$, the estimated eigenvectors, to a rotated version of $\tilde{\Psi}$, the true eigenvectors. This is essentially bounding the deviation in the principal subspace after perturbing a matrix with entry-wise independent random noise. The traditional tool for such a bound is the Davis-Kahan $\sin\Theta$ Theorem. Here we use a different result, first given in [22, 23], to obtain a bound that is better suited to our task.

Lemma 5 (Accuracy of principal subspace)**.** *Let $\hat{U} \in \mathbb{R}^{n \times K}$ be the K leading eigenvectors of A. There exists a $K \times K$ orthogonal matrix Q such that*

$$\|\hat{U} - \tilde{\Psi}Q\|_F \leq \frac{2\sqrt{2K}}{\alpha_n \lambda_n n_{\min}}\|A - P\|.$$

The proof is adapted from Theorem 3.3 of [23], which uses a novel lower bound on curvature of the PCA objective function at the principal subspace.

The second major step is to control the spectral norm of the noise matrix $A - P$. Common techniques include the matrix Bernstein inequality [21], and combinatorial arguments [9, 16]. The following lemma combines the results given by different methods under different conditions.

Lemma 6 (Spectral norm bound of noise matrix)**.** *For any $a, r > 0$, there exists a constant $c(a, r)$ depending only on a and r such that with probability at least $1 - 2n^{-r}$ we have,*

$$\|A - P\| \leq \begin{cases} c(a, r)(K \wedge \sqrt{\log n})\sqrt{n\alpha_n}, & \text{if } \alpha_n \geq a\frac{\log n}{n}, \\ c(a, r)\sqrt{n\alpha_n}, & \text{if } \alpha_n \geq a\frac{(\log n)^4}{n}. \end{cases} \tag{6}$$

The constant $c(a, r)$ in Lemma 6 is closely related to the $c_0(a, r, \epsilon)$ in Theorem 3. Its closed form is given in the proof of Lemma 6 in Supplementary Material.

Combining Lemmas 5 and 6, we know that $\hat{U}$ is close to $\tilde{\Psi}Q$ in Frobenius norm for some $K \times K$ orthonormal Q. The following lemma controls the error rate of the (approximate) k-means solution in terms of $\|\hat{U} - \tilde{\Psi}Q\|_F$. It is a refinement of the arguments used in Theorem 3.1 of [20] and Theorem 2.2 of [12].

Lemma 7 (Hamming error of k-means solution)**.** *For $\epsilon > 0$ and any two matrices $\hat{V}, V$ with same dimension such that $V = \Psi X$ with $\Psi \in \mathbb{M}^{n\times K}$, $X \in \mathbb{R}^{K\times K}$, let $\bar{V} = \hat{\Psi}\hat{X}$ be a $(1+\epsilon)$-approximate solution to the k-means problem in eq.* (2) *with $\hat{U}$ replaced by $\hat{V}$. Define $\delta_k = \min_{\ell\neq k} \|X_{\ell\cdot} - X_{k\cdot}\|$ and $S_k = \{i \in G_k : \|\bar{V}_{i\cdot} - V_{i\cdot}\| \geq \delta_k/2\}$ then*

$$\sum_{k=1}^{K} |S_k|\delta_k^2 \leq 4(4+2\epsilon)\|\hat{V} - V\|_F^2. \tag{7}$$

Moreover, if $(16+8\epsilon)\|\hat{V} - V\|_F^2/\delta_k^2 < |G_k|$ for all k, then there exists a $K \times K$ permutation matrix J such that $\hat{\Psi}_{S\cdot} = \Psi_{S\cdot}J$, where $S = \cup_{k=1}^{K}(G_k \backslash S_k)$.

Lemma 7 suggests that the (approximate) k-means solution gives correct clustering for all nodes in the set S. Thus all mis-clustered nodes are contained in S^c. The cardinality of S^c can be bounded by $\|\hat{V} - V\|_F^2/\delta_k^2$. In the proof of Theorem 3, Lemma 7 is applied to $\hat{U}$ and $\tilde{\Psi}Q$ for some orthonormal Q, and $\|\hat{U} - \tilde{\Psi}Q\|_F^2$ is bounded by Lemmas 5 and 6.

3.2 Application to Planted Partition Models

The planted partition model is a special case of the stochastic block model, where the within-block edge probability is $p \in (0, 1]$ and between-block edge probability is $q \in (0, p)$. Equivalently, the block connectivity matrix B can be written as $(p-q)I_K + q\mathbf{1}\mathbf{1}^T$. In our scaling (3), this corresponds to the following parametrization.

$$p = \alpha_n, \quad q = \alpha_n(1-\lambda_n), \quad B_0 = \lambda_n I_K + (1-\lambda_n)\mathbf{1}\mathbf{1}^T.$$

In this case λ_n is indeed the smallest absolute eigenvalue of B_0. Then Theorem 3 implies that a sufficient condition for the spectral clustering method to be overall consistent is

$$p = \Omega\left(\frac{\log n}{n}\right), \quad \frac{p-q}{\sqrt{p}} = \begin{cases} \Omega\left(\sqrt{K}(K \wedge \sqrt{\log n})\frac{\sqrt{n}}{n_{\min}}\right) \\ \omega\left(\sqrt{K}(K \wedge \sqrt{\log n})\frac{\sqrt{n_{\max}}}{n_{\min}}\right) \end{cases},$$

where the first part corresponds to the overall sparsity condition and the second part corresponds to the block separation (the normalized difference between within and between block connectivity).

The planted partition model offers a natural benchmark for the performance of community recovery algorithms. The existing results (see , e.g., [7, 6]) typically exhibit a lower bound on $(p-q)/\sqrt{p}$ as the main sufficient condition for consistent community detection. However, the condition on the overall sparsity p, although playing a very important role, is less commonly recognized.

To further interpret our result and compare it with the state-of-art methods, we consider two special cases of the planted partition models. The comparison indicates that the conditions required by our method are weaker when the number of blocks is small. We note that the comparison itself is not sufficient to claim that one method is superior since different notions of consistency are used. However, out algorithm has the advantage of being computationally less demanding.

3.2.1 The case of $K = O(1)$.

This corresponds to a constant number of blocks. We have $K \wedge \sqrt{\log n} \leq K = O(1)$. Our separation condition becomes

$$\frac{p-q}{\sqrt{p}} = \omega\left(\frac{\sqrt{n}}{n_{\min}}\right), \tag{8}$$

and the comparing conditions are

$$\frac{p-q}{\sqrt{p}} = \Omega\left(\frac{\sqrt{n}\log^2 n}{n_{\min}}\right) \tag{9}$$

in [7], and

$$\frac{p-q}{\sqrt{p}} = \Omega\left(\frac{\sqrt{n\log n}}{n_{\min}}\right) \tag{10}$$

in [6] (provided that $q = \Omega(1)$).

In the case of $K = O(1)$, it is of interest to see how large $n_{\min}$ needs to be for the methods to achieve consistency. Eq. (8) can hold when $n_{\min} = \omega(\sqrt{n})$, while eqs. (9) and (10) require $n_{\min} = \Omega(\sqrt{n}\log^2 n)$, and $n_{\min} = \Omega(\sqrt{n\log n})$, respectively.

3.2.2 The case of $K = \omega(1)$ and $n_{\max} = n_{\min} = n/K$.

When the number of blocks diverges as n increases, the interaction between K, $n_{\min}$, $n_{\max}$ and λ_n becomes more complicated. We consider a further special case where all clusters have the same size: $n_{\max} = n_{\min} = n/K$. Now our separation condition becomes

$$\frac{p-q}{\sqrt{p}} = \Omega\left(\frac{K^{3/2}(K\wedge\sqrt{\log n})}{\sqrt{n}}\right) . \tag{11}$$

In comparison, the sufficient condition in [7] is

$$\frac{p-q}{\sqrt{p}} = \Omega\left(\frac{K\log^2 n}{\sqrt{n}}\right) , \tag{12}$$

which is stronger than (11) when $K = o(\log^3 n)$. For another comparison, the sufficient condition given in [6] is, provided that $q = \Omega(1)$,

$$\frac{p-q}{\sqrt{p}} = \Omega\left(\frac{K\sqrt{\log n}}{\sqrt{n}}\right) , \tag{13}$$

which is weaker than (12). But the additional assumption $q = \Omega(1)$ implies that $\alpha_n = \Omega(1) = \Omega(\log^4 n/n)$ so (11) can be reduced to the even weaker condition: $\frac{p-q}{\sqrt{p}} = \Omega(K^{3/2}/\sqrt{n})$, which is weaker than (13) when $K = o(\log n)$.

The overall sparsity is not explicitly considered in [7]. However, it is easy to check that (12) cannot hold when $p = a\log n/n$ for some positive constant a. While in this case our condition (11) can still hold as long as

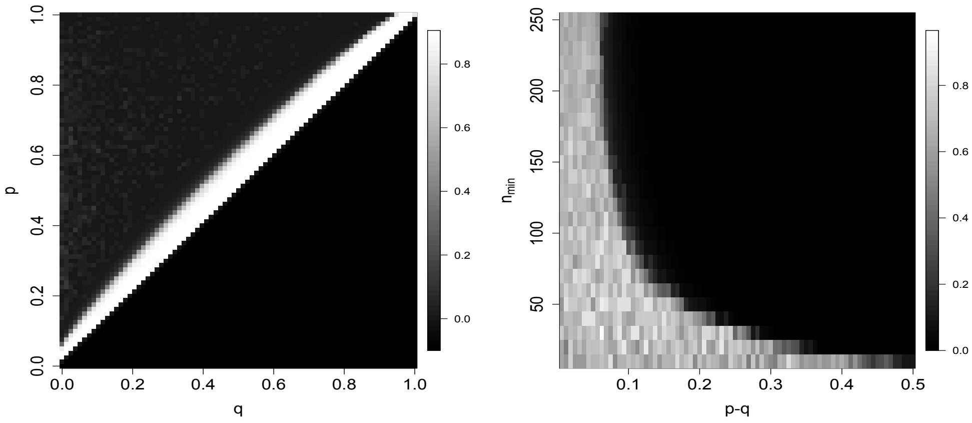
$$\lambda_n = \omega\left(\frac{K^{3/2}(K\wedge\sqrt{\log n})}{\sqrt{\log n}}\right) ,$$

which is possible if $\lambda_n = \Omega(1)$ and $K = o\left((\log n)^{1/4}\right)$.

4 Numerical Examples

In this section we describe the results of two simulation experiments in order to illustrate the dependence of the performance of the spectral algorithm proposed here on some of the key parameters used in our analysis.

In the first experiment we consider the same setting used in the simulation study described in [7]: we generated random graphs according to a planted partition model with $K = 5$ communities of equal size 200 and over a regular grid of values for the parameters $p \geq q$ (we set the side length of each square in the grid to 1/80). For each choice of the pair (p, q) in the grid we computed the average Hamming error rate $\mathrm{Err}_n(\hat{\Psi}, \Psi)$ given by the proposed algorithm over 10 simulations of the model. We used the k-means++ subroutine ([3]) to solve the optimization problem (2). For every simulation we used 4 different starting values and then took the solution giving the minimal Hamming error rate. Plot (a) in Figure 1 shows the results of the experiment. For each value of p and q, with $p \geq q$, the color of the corresponding square in the grid indicates the average value of the normalized Hamming distance over the simulations, with darker colors corresponding to small values and lighter color to larger values. For $p < q$ we set the value to 0. From the figure it is clear that the smaller the difference $p - q$ the larger the error rate and the worse the performance of the algorithm, as expected. We can also see that the degradation in the performance is not uniform



(a) (b)

Figure 1: Average Hamming error rate over 10 simulations from planted partition models. Left: $n = 1000$, $K = 5$, $n_{\min} = n_{\max} = 200$. Right: $n = 500$, $K = 2$.

over p and q and appears to be worse when both p and q are close to 0.5. Finally, there appears to be a phase transition in the performance of the algorithm as a function of $p - q$ and q: as soon as the difference $p - q$ crosses a certain value (which seems to depend on q and is maximal around $q = 0.5$), the algorithm transitions from having very poor performance to working almost perfectly. When we visually compare these results with the results from the same experiment based on the algorithm proposed in [7] (as shown in Figure 1 in that paper), we see that our simpler and faster algorithm seems to behave very similarly (though, to be fair, in our experiment we use the Hamming error rate, which gives only approximate recovery, while [7] measures perfect recovery).

In our second experiment, we simulated from a planted partition model with $n = 500$ and $K = 2$. Here we fix $q = 0.5$ but let p vary from 0.5 to 1 in steps of equal size $1/80$ and let $n_{\min}$ vary 10 to 250 in increments of 10. Just like in the previous experiment, for every choice of (p, q) and $n_{\min}$, we simulated the model 10 times and took the average Hamming error rate $\mathrm{Err}_n(\hat{\Psi}, \Psi)$ based on the best result out of 4 runs of the k-means++ algorithm. Plot (b) in Figure 1 displays the results using on the same coloring scheme of the previous experiment. It can be seen that, when p and q are very close to each other, larger values of $n_{\min}$ will increase the performance of the algorithm only marginally. But when the gap between p and q widens, small increases in the value of $n_{\min}$ will greatly boost the performance, as predicted by our analysis. Interestingly, like in the previous experiment, there also appears to be a phase transition as a function of $n_{\min}$ and $p - q$.

5 Discussion

In this work we study the performance of spectral clustering for general stochastic block models. We have shown that a simple and practical spectral clustering algorithm can give good community detection even for very sparse models. When the number of blocks is relatively small, the condition required for (approximate) consistency is weaker than most existing works. The method and argument presented in this paper are general enough so that they can be refined and/or extended to other situations. For example, we believe that similar analysis can be carried out for spectral clustering using graph Laplacians, such as in [20, 6]. On the other hand, one can easily combine the arguments given in this paper and [12] to obtain spectral clustering error bounds for degree corrected block models. Moreover, one may conjecture that the $(K \wedge \sqrt{\log n})$ term in our error bound may be replaced by 1 without the additional assumption of $\alpha_n = \Omega(\log^4 n/n)$. This would require a finer spectral bound of random matrices, perhaps using a different technique. These extensions and refinements, together with the phase transitions observed in the simulation study, are all interesting questions and will be pursued in future work.

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

[2] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor spectral approach to learning mixed membership community models. *arXiv preprint arXiv:1302.2684*, 2013.

[3] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithm (SODA)*, 1027–1035, 2007.

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

[5] A. Celisse, J.J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6, 1847–1899, 2012.

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

[7] Y. Chen, S. Sanghavi, and H. Xu. Clustering sparse graphs. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2213–2221. 2012.

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

[9] A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19:227–284, 2010.

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

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

[12] J. Jin. Fast community detection by SCORE. arXiv:1211.5803, 2012.

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

[14] E. D. Kolaczyk. *Statistical analysis of network data*. Springer, 2009.

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

[16] L. Lu and X. Peng. Spectra of edge-independent random graphs. *arXiv preprint arXiv:1204.6207*, 2012.

[17] F. McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

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

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

[20] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39:1878–1915, 2011.

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

[22] V. Vu and J. Lei. Minimax rates of estimation for sparse pca in high dimensions. *JMLR: Workshop and Conference Proceedings*, 22:1278–1286, 2012.

[23] V. Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *arXiv preprint arXiv:1211.0373*, 2012.

Supplementary Material

Keeping track of the constant function $c_0(a, r, \epsilon)$

When $\alpha_n \geq a \log n/n$, we have

$$\begin{aligned} c_0(a,r,\epsilon) =& 64(2+\epsilon)\left[\bar{c}_1(a,r) \vee \bar{c}_2(a,r)\right]^2 \,, & (14)\\ \bar{c}_1(a,r) =& c^* \left[2 + \frac{1+r+\sqrt{(1+r)^2+18a(1+r)}}{3a}\right]^{1/2} \,, & (15)\\ \bar{c}_2(a,r) =& 1 + \frac{1+r}{3\sqrt{a}} + \sqrt{\frac{(1+r)^2}{9a} + 2(1+r)} \,, & (16) \end{aligned}$$

where c^* is a universal constant that appears in an combinatorial upper bound of the spectral norm of $A - P$ ([9]).

The case of $\alpha_n \geq a \log^4 n/n$ is very similar and the constant $c_0(a, r, \epsilon)$ can be recovered from the second part of the proof of Lemma 6 below.

Technical proofs

Proof of Theorem 3. We will only prove the case $\alpha_n \geq a \log n/n$. The other case $\alpha_n \geq a(\log n)^4/n$ can be treated similarly by using a difference spectral bound of $\|A - P\|$.

Define

$$c(a,r) \equiv 8 \max\left[(c^*)^2(1 + c_1(a,r)), (1 + c_2(a,r))^2\right] \,,$$

where c_1 and c_2 are defined in (21), (22), and c^* the universal constant in the proof of Lemma 6.

By Lemmas 5 and 6 we have, for some orthogonal matrix Q and with probability at least $1 - 2n^{-r}$,

$$\|\hat{U} - \tilde{\Psi}Q\|_F^2 \leq \frac{c(a,r)K(K \wedge \sqrt{\log n})^2}{\alpha_n \lambda_n^2} \frac{n}{n_{\min}^2} \,.$$

Now apply Lemma 7 to $\hat{V} = \hat{U}$ and $V = \tilde{\Psi}Q = \Psi X$ where $X = D^{-1}Q$ with $D = \mathrm{diag}(\sqrt{n_1}, \sqrt{n_2}, ..., \sqrt{n_K})$. Denote the $(1+\epsilon)$-approximate solution by $\hat{\Psi}\hat{X}$ where $\hat{\Psi} \in \mathbb{M}^{n \times K}$ and $\hat{X} \in \mathbb{R}^{K \times K}$.

Recall that in Lemma 7 we define $\delta_k = \min_{\ell \neq k} \|X_{\ell \cdot} - X_{k \cdot}\|$. In this case, $\delta_k^2 = n_k^{-1} + \min_{\ell \neq k} n_\ell^{-1} \geq n_k^{-1}$ because the rows of Q have unit norm and are orthogonal to each other.

Also recall that $S_k = \{i \in G_k : \|\bar{V}_{i\cdot} - V_{i\cdot}\| \geq \delta_k/2\}$. Then by (7) in Lemma 7,

$$\sum_{k=1}^K \frac{|S_k|}{n_k} \leq \sum_{k=1}^K |S_k|\delta_k^2 \leq 8(2+\epsilon)\|\hat{V} - V\|_F^2 \leq \frac{c_0(a,r,\epsilon)K(K \wedge \sqrt{\log n})^2}{\alpha_n \lambda_n^2} \frac{n}{n_{\min}^2} \,, \tag{17}$$

where

$$c_0(a,r,\epsilon) \equiv 8(2+\epsilon)c(a,r) \,.$$

According to Lemma 7, one can find a permutation matrix J such that $\hat{\Psi}_{S\cdot} = \Psi_{S\cdot}J$, where $S = \cup_{k=1}^K (G_k \backslash S_k)$. Then with probability at least $1 - 2n^{-r}$,

$$\|\hat{\Psi} - \Psi J\|_F^2 = \|\hat{\Psi}_{S^c\cdot} - \Psi_{S^c\cdot}J\|_F^2 \leq \sum_{k=1}^K 2|S_k| \leq \frac{2c_0(a,r,\epsilon)K(K \wedge \sqrt{\log n})^2 n_{\max} n}{\alpha_n \lambda_n^2 n_{\min}^2} \,. \quad \square$$

Proof of Lemma 5. First we assume that B is positive semidefinite and hence so is P. Write $\hat{U} = (\hat{u}_1, ..., \hat{u}_K)$ in the following equivalent optimization formulation.

$$\begin{aligned} \hat{U} =& \arg\max_U \left\langle A, UU^T \right\rangle & (18)\\ &\text{s.t. } U \in \mathbb{R}^{n \times K}, \ \ U^T U = I_K . & (19) \end{aligned}$$

where $\langle X, Y\rangle = \mathrm{trace}(X^T Y)$ for matrices X, Y with compatible dimensions.

Then by a result on the curvature of PCA optimization problem (18) at the principal subspace (Lemma 4.2 and Proposition 2.2 of [23]) we have

$$\|\hat{U}\hat{U}^T - \tilde{\Psi}\tilde{\Psi}^T\|_F^2 \leq \frac{2}{\alpha_n \lambda_n n_{\min}} \left\langle A - P, \hat{U}\hat{U}^T - \tilde{\Psi}\tilde{\Psi}^T \right\rangle . \tag{20}$$

Now let $\Upsilon = \frac{\hat{U}\hat{U}^T - \tilde{\Psi}\tilde{\Psi}^T}{\|\hat{U}\hat{U}^T - \tilde{\Psi}\tilde{\Psi}^T\|_F}$. Then $\|\Upsilon\|_F = 1$. By Theorem I.5.5 of [26], we have

$$\Upsilon = \sum_{\ell=1}^{K} a_\ell (\eta_\ell \eta_\ell^T - \tau_\ell \tau_\ell^T),$$

where a_ℓ are positive numbers satisfying $\sum_{\ell=1}^{K} a_\ell^2 = 1/2$ and $(\eta_1, \tau_1, \eta_2, \tau_2, ..., \eta_K, \tau_K)$ are orthonormal vectors in $\mathbb{R}^n$. Then

$$\langle A - P, \Upsilon\rangle = \sum_{\ell=1}^{K} a_\ell \left[\eta_\ell^T (A-P)\eta_\ell - \tau_\ell^T (A-P)\tau_\ell\right] \leq \sqrt{2K}\|A - P\|.$$

Combining with (20) we have

$$\|\hat{U}\hat{U}^T - \tilde{\Psi}\tilde{\Psi}^T\|_F \leq \frac{2\sqrt{2K}}{\alpha_n \lambda_n n_{\min}} \|A - P\| .$$

The desired result follows because by proposition 2.2 of [23], there exists a K dimensional orthogonal matrix Q such that

$$\|\hat{U} - \tilde{\Psi}Q\|_F \leq \|\hat{U}\hat{U}^T - \tilde{\Psi}\tilde{\Psi}^T\|_F.$$

Now consider general matrix B. For any matrix X, let $X_{\mathrm{aug}} = \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix}$ be the augmented matrix.

According to Lemma 8 and the definition of $\hat{U}$, we have

$$\hat{U}_{\mathrm{aug}} \equiv \begin{pmatrix} \hat{U}/\sqrt{2} \\ \hat{U}\Sigma_K/\sqrt{2} \end{pmatrix} = \arg\max_{U} \langle A_{\mathrm{aug}}, UU^T\rangle$$
$$\text{s.t. } U \in \mathbb{R}^{2n\times K},\ \ U^T U = I_K ,$$

where Σ_K is the $K \times K$ diagonal matrix whose diagonal entries correspond to the signs of the kth eigenvalue of A, and $\langle X, Y\rangle = \mathrm{trace}(X^T Y)$ for matrices X, Y with compatible dimensions. A similar relationship holds for $\tilde{\Psi}$, P_{aug}, with matrices $\tilde{\Psi}_{\mathrm{aug}}$ and Σ_K^* in obvious correspondence.

Then by the same argument we have

$$\begin{aligned} &\|\hat{U}_{\mathrm{aug}}\hat{U}_{\mathrm{aug}}^T - \tilde{\Psi}_{\mathrm{aug}}\tilde{\Psi}_{\mathrm{aug}}^T\|_F^2 \\ \leq &\frac{2}{\alpha_n \lambda_n n_{\min}} \left\langle A_{\mathrm{aug}} - P_{\mathrm{aug}}, \hat{U}_{\mathrm{aug}}\hat{U}_{\mathrm{aug}}^T - \tilde{\Psi}_{\mathrm{aug}}\tilde{\Psi}_{\mathrm{aug}}^T \right\rangle . \end{aligned}$$

the rest of the proof follows the case of positive definite B and the following fact (Corollary VII.5.6 of [24])

$$\|\hat{U} - \tilde{\Psi}\|_F^2 \leq \frac{1}{2}\|\hat{U}_{\mathrm{aug}} - \tilde{\Psi}_{\mathrm{aug}}\|_F^2 . \qquad \square$$

The following elementary result, which can be directly verified, gives the eigen structure of X_{aug}.

Lemma 8. *Let $X = UDV^T$ be the singular value decomposition of X. Then X_{aug} has eigen decomposition $X_{\mathrm{aug}} = \begin{pmatrix} U/\sqrt{2} & U/\sqrt{2} \\ V/\sqrt{2} & -V/\sqrt{2} \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & -D \end{pmatrix} \begin{pmatrix} U/\sqrt{2} & U/\sqrt{2} \\ V/\sqrt{2} & -V/\sqrt{2} \end{pmatrix}$.*

Proof of Lemma 6. We prove two cases separately.

The case of $\alpha_n \geq a \log n/n$**.** Let $s_j = \sum_{i \neq j} A_{ij}$ be the degree of node j. Remember that $\max_{1 \leq k, \ell \leq K} b_0(k, \ell) = 1$. Because $\alpha_n \geq a \log n/n$, a standard application of Bernstein's inequality and union bound imply that, for all $c > 1$,

$$\begin{aligned}\Pr\left(\max_{1\leq j\leq n} s_j \geq c\alpha_n n\right) \leq & \Pr\left(\max_{1\leq j\leq n} s_j - \mathbb{E}s_j \geq (c-1)\alpha_n n\right) \\ \leq & n \exp\left(-\frac{\frac{1}{2}(c-1)^2\alpha_n^2 n^2}{\alpha_n n + \frac{1}{3}(c-1)\alpha_n n}\right) \\ = & n \times n^{-\frac{3(c-1)^2}{2c-4}\frac{\alpha_n n}{\log n}} \\ \leq & n \times n^{-\frac{3(c-1)^2}{2c-4}a} \leq n^{-r}\,,\end{aligned}$$

where the last inequality holds for all $c \geq c_1(a, r)$ with

$$c_1(a,r) = 1 + \frac{1 + r + \sqrt{(1+r)^2 + 18a(1+r)}}{3a}\,. \tag{21}$$

Applying Lemma 8.5 of [9], we have for a universal constant c^*,

$$\Pr\left(\|A - P\| \geq c^* K\sqrt{(1+c)}\sqrt{\alpha_n n}\right) \leq n^{-r}.$$

Next we use matrix Bernstein inequality to prove that $\Pr(\|A - P\| \geq c\sqrt{n\alpha_n \log n}) \leq n^{-r}$ for $c \geq 1 + c_2(a, r)$, where

$$c_2(a,r) = \frac{1+r}{3\sqrt{a}} + \sqrt{\frac{(1+r)^2}{9a} + 2(1+r)}\,. \tag{22}$$

First notice that, since $P = \mathbb{E}[A] + \mathrm{diag}(P)$, by the triangle inequality,

$$\|A - P\| = \|A - \mathbb{E}[A] + \mathrm{diag}(P)\| \leq \|A - \mathbb{E}[A]\| + \max_i p_{i,i} \leq \|A - \mathbb{E}[A]\| + \alpha_n.$$

We will bound the first term. To that end, for $i, j \in [n]$, we denote with $E^{(i,j)}$ the $n \times n$ matrix with all entries set to 0 except for the (i, j)th entry, which is 1. Then,

$$A - \mathbb{E}[A] = \sum_{i<j} \overline{A}^{(i,j)} \equiv \sum_{i<j} (a_{i,j} - p_{i,j})(E^{(i,j)} + E^{(j,i)}),$$

where we recall that the random variables $a_{i,j} \sim \mathrm{Bernoulli}(p_{i,j})$ are mutually independent, for $i < j$. Next, it is immediate to see that $\|\overline{A}^{(i,j)}\| \leq 1$ and that $\mathbb{E}[(\overline{A}^{(i,j)})^2] = p_{i,j}(1-p_{i,j})(E^{(i,i)}+E^{(j,j)})$. Since $p_{i,j}(1 - p_{i,j}) \leq \alpha_n$, we obtain that

$$\left\|\sum_{i<j} \mathbb{E}[(\overline{A}^{(i,j)})^2]\right\| \leq \alpha_n n\,.$$

Then, using the matrix Bernstein inequality (Theorem 1.4 of [21], see also [25]), we have

$$\Pr\left(\|A - \mathbb{E}[A]\| \geq t\right) \leq n \exp\left\{-\frac{t^2/2}{n\alpha_n + t/3}\right\}. \tag{23}$$

By assumption, $\alpha_n \geq a\frac{\log n}{n}$. Thus, setting $t = c\sqrt{n\alpha_n \log n}$, (23) yields

$$\Pr\left(\|A - \mathbb{E}[A]\| \geq c\sqrt{\alpha_n n \log n}\right) \leq \frac{1}{n^r}\,,$$

for $c \geq c_1(a, r)$, where $c_2(a, r)$ is defined as in (22). Thus we have $\Pr(\|A - P\| \geq c\sqrt{n\alpha_n \log n} \leq n^{-r}) \leq n^{-r}$ for all $c \geq 1 + c_2(a, r)$.

The case of $\alpha_n \geq a \log^\gamma n/n$, for $\gamma \geq 4$. This case can be handled using directly the results of [16]. Write, for simplicity, $\Delta_n = \alpha_n n$. By Theorem 6 in [16] (whereby in the notation of that paper we let $K = 1$), for any $C > 0$ and $\lambda \leq \left(\frac{\Delta_n}{32}\right)^{1/4}$,

$$\Pr\Big(\|A - P\| \geq 2\sqrt{\Delta_n} + C\Delta_n^{1/4}\log n\Big) \leq 4n \exp\left\{-\lambda \log\left(1 + \frac{C}{2}\Delta_n^{-1/4}\log n\right)\right\}, \tag{24}$$

Let $c = 32^{1/4}(r+2) \geq 32^{1/4}\left(r + 1 + \frac{\log 4}{\log n}\right)$ under the assumption that $n \geq 4$. Since $\Delta_n \geq a(\log n)^4$ for some $a > 0$, setting $C = 2\frac{e^{ca^{-1/4}}-1}{a^{-1/4}}$ will guarantee that $\log\left(1 + \frac{C}{2}\Delta_n^{-1/4}\log n\right) \geq \Delta_n^{-1/4} c \log n$, for all n. Then, following closely the arguments in Lemma 1 and Theorem 6 of [16], choosing $\lambda = \left(\frac{\Delta_n}{32}\right)^{1/4}$, yields by (24) that

$$\Pr\Big(\|A - P\| \geq C'\sqrt{\alpha_n n}\Big) \leq \Pr\Big(\|A - P\| \geq 2\sqrt{\Delta_n} + C\Delta_n^{1/4}\log n\Big) \leq \frac{1}{n^r},$$

where $C' = 2 + Ca^{-1/4}$. The case of $\gamma \geq 4$ follows trivially. □

Proof of Lemma 7. First by the definition of $\bar{V}$ and the fact that V is feasible for problem (2) we have $\|\bar{V} - V\|_F^2 \leq 2\|\bar{V} - \hat{V}\|_F^2 + 2\|\hat{V} - V\|_F^2 \leq (4 + 2\epsilon)\|\hat{V} - V\|_F^2$. Let $S_k = \{i \in G_k : \|\bar{V}_{i\cdot} - V_{i\cdot}\| \geq \delta_k/2\}$. Then

$$\sum_{k=1}^{K} |S_k|\delta_k^2/4 \leq \|\bar{V} - V\|_F^2 \leq (4 + 2\epsilon)\|\hat{V} - V\|_F^2\,. \tag{25}$$

which concludes the first claim of the lemma.

Equation (25) also implies that

$$|S_k| \leq (16 + 8\epsilon)\|\hat{V} - V\|_F^2/\delta_k^2 < |G_k|, \ \ \forall\, k.$$

Therefore $T_k \equiv G_k \backslash S_k \neq \emptyset$. If $i \in T_k$ and $j \in T_\ell$ with $k \neq \ell$, then $\bar{V}_{i\cdot} \neq \bar{V}_{j\cdot}$ because otherwise $\|V_{i\cdot} - V_{j\cdot}\| \leq \|V_{i\cdot} - \bar{V}_{i\cdot}\| + \|V_{j\cdot} - \bar{V}_{j\cdot}\| < \delta_k/2 + \delta_\ell/2$, contradicting with the definition of δ_k and δ_ℓ. On the other hand, $\bar{V}$ has at most K distinct rows. As a result, we must have $\bar{V}_{i\cdot} = \bar{V}_{j\cdot}$ if $i, j \in T_k$ for some k, and $\bar{V}_{i\cdot} \neq \bar{V}_{j\cdot}$ if $i \in T_k$, $j \in T_\ell$ with $k \neq \ell$. This gives a correspondence of clustering between the rows in $\bar{V}_{S\cdot}$ and those in $V_{S\cdot}$ where $S = \cup_{k=1}^{K} T_k$. □

Additional References for Supplementary Material

[24] R. Bhatia. *Matrix analysis*, volume 169. Springer, 1997.

[25] F. Chung and M. Radcliffe. On the spectra of general random graphs. *the electronic journal of combinatorics*, 18(P215):1, 2011.

[26] G. W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.