Analysis of spectral clustering algorithms for community detection: the general bipartite setting

Zhixin Zhou   ZHIXIN@G.UCLA.EDU

Department of Statistics
University of California
Los Angeles, CA 90095-1554, USA

Arash A. Amini   AAAMINI@UCLA.EDU

Department of Statistics
University of California
Los Angeles, CA 90095-1554, USA

Editor: Ulrike von Luxburg

Abstract

We consider spectral clustering algorithms for community detection under a general bipartite stochastic block model (SBM). A modern spectral clustering algorithm consists of three steps: (1) regularization of an appropriate adjacency or Laplacian matrix (2) a form of spectral truncation and (3) a k-means type algorithm in the reduced spectral domain. We focus on the adjacency-based spectral clustering and for the first step, propose a new data-driven regularization that can restore the concentration of the adjacency matrix even for the sparse networks. This result is based on recent work on regularization of random binary matrices, but avoids using unknown population level parameters, and instead estimates the necessary quantities from the data. We also propose and study a novel variation of the spectral truncation step and show how this variation changes the nature of the misclassification rate in a general SBM. We then show how the consistency results can be extended to models beyond SBMs, such as inhomogeneous random graph models with approximate clusters, including a graphon clustering problem, as well as general sub-Gaussian biclustering. A theme of the paper is providing a better understanding of the analysis of spectral methods for community detection and establishing consistency results, under fairly general clustering models and for a wide regime of degree growths, including sparse cases where the average expected degree grows arbitrarily slowly.

Keywords: Spectral clustering; bipartite networks; stochastic block model; regularization of random graphs; community detection; sub-Gaussian biclustering; graphon clustering.

1. Introduction

Spectral clustering is one of the most popular and successful approaches to clustering and has appeared in various contexts in statistics and computer science among other disciplines. The idea generally applies when one can define a similarity matrix between pairs of objects to be clustered (Ng et al., 2002; Von Luxburg, 2007; Von Luxburg et al., 2008). Recently, there has been a resurgence of interest in the spectral approaches and their analysis in the context of network clustering (Boppana, 1987; McSherry, 2001; Dasgupta et al., 2006;

Coja-Oghlan, 2010; Tomozei and Massoulié, 2010; Rohe et al., 2011; Balakrishnan et al., 2011; Chaudhuri et al., 2012; Fishkind et al., 2013; Qin and Rohe, 2013; Joseph and Yu, 2013; Krzakala et al., 2013; Lei et al., 2015; Yun and Proutiere, 2014a,b; Binkiewicz et al., 2014; Lyzinski et al., 2014; Chin et al., 2015; Gao et al., 2017). The interest has been partly fueled by the recent activity in understanding statistical network models for clustering and community detection, and in particular by the flurry of theoretical work on the stochastic block model (SBM)—also called the planted partition model—and its variants (Abbe, 2017).

There has been significant recent advances in analyzing spectral clustering approaches in SBMs. We start by identifying the main components of the analysis in Section 3, and then show how variations can be introduced at each step to obtain improved consistency results and novel algorithms. We will work in the general bipartite setting which has received comparatively less attention, but most results in the paper can be easily extended to the (symmetric) unipartite models (cf. Remark 7). It is worth noting that clustering bipartite SBMs is closely related to the biclustering (Hartigan, 1972) and co-clustering (Dhillon, 2001; Rohe and Yu, 2012) problems.

A modern spectral clustering algorithm often consists of three steps: (a) the regularization and concentration of the adjacency matrix (or the Laplacian) (b) the spectral truncation step, and (c) the k-means step. By using variations in each step one obtains different spectral algorithms, which is then reflected in the variations in the consistency results.

The regularization step is fairly recent and is motivated by the observation that proper regularization significantly improves the performance of spectral methods in sparse networks (Chaudhuri et al., 2012; Amini et al., 2013; Joseph and Yu, 2013; Le et al., 2015; Chin et al., 2015). In particular, regularization restores the concentration of the adjacency matrix (or the Laplacian) around its expectation in the *sparse regime*, where the average degree of the network is constant or grows very slowly with the number of nodes. In this paper, building on these recent advances, we introduce a novel regularization scheme for the adjacency matrix that is fully data-driven and avoids relying on unknown quantities such as maximum expected degrees (for rows and columns) of the network. This regularization scheme is introduced as Algorithm 1 in Section 4 and we show that under a general SBM it achieves the same concentration bound (Theorem 3) as its oracle counterpart.

For the spectral truncation step, we will consider three variations, one of which (Algorithm 2) is the common approach of keeping the top k leading eigenvectors as columns of the matrix passed to the k-means step. In this traditional approach, the spectral truncation can be viewed as obtaining a low-dimensional representation of the data, suitable for an application of simple k-means type algorithms. We also consider a recent variant (Algorithm 3) proposed in Yun and Proutiere (2014b) and Gao et al. (2017), in which the spectral truncation step acts more as a denoising step. We then propose a third alternative (Algorithm 4 in Section 5) which combines the benefits of both approaches while improving the computational complexity. One of our novel contributions is to derive consistency results for Algorithms 3 and 4, under a general SBM, showing that the behavior of the two algorithms is the same (but different than Algorithm 2) assuming that the k-means step satisfies a property we refer to as *isometry invariance* (Theorems 5 and 6).

In the final step of spectral clustering, one runs a k-means type algorithm on the output of the truncation step. We discuss this step in some detail since it is often mentioned

briefly in the analyses of spectral clustering, with the exception of a few recent works (Lei et al., 2015; Gao et al., 2017, 2018). By the k-means step, we do not necessarily mean the solution of the well-known k-means problem, although, this step is usually implemented by an algorithm that approximately minimizes the k-means objective. We will consider the k-means step in some generality, by introducing the notation of a *k-means matrix* (cf. Section 3.3). The goal of the final step of spectral clustering is to obtain a k-means matrix which is close to the output of the truncation step. We characterize sufficient conditions on this approximation so that the overall algorithm produces consistent clustering. Any approach that satisfies these conditions can be used in the k-means step, even if it is not an approximate k-means solver.

Most of the above ideas extend beyond SBMs and, in Section 6, we consider various extensions. We first consider some unique aspects of the bipartite setting, for example, the possibility of having clusters only on one side of the network, or having more clusters on one side than the rank of the expected adjacency matrix. We then show how the results extend to general sub-Gaussian biclustering (Section 6.3) and general inhomogeneous random graphs with approximate clusters (Section 6.4).

The organization of the rest of the paper is as follows: After introducing some notation in Section 1.1, we discuss the general (bipartite) SBM in Section 2. An outline of the analysis is given in Section 3, which also provides the high-level intuition of why the spectral clustering works in SBMs and what each step will achieve in terms of guarantees on its output. This section provides a typical blueprint theorem on consistency (Theorem 1 in Section 3.1) which serves as a prelude to later consistency results. The rest of Section 3 provides the details of the last two steps of the analysis sketched in Section 3.1. The regularization and concentration (the first step) is detailed in Section 4 where we also introduce our data-driven regularization. We then give explicit algorithms and their corresponding consistency results in Section 5. Extensions of the results are discussed in Section 6. Some simulations showing the effectiveness of the regularization are provided in Section 7.

Related work.   There are numerous papers discussing aspect of spectral clustering and its analysis. Our paper is mostly inspired by recent developments in the field, especially by the consistency results of Lei et al. (2015), Yun and Proutiere (2014a), Chin et al. (2015), Gao et al. (2017) and concentration results for the regularized adjacency (and Laplacian) matrices such as Le et al. (2015) and Bandeira et al. (2016). Theorem 4 on the consistency of the typical adjacency-based spectral clustering algorithm—which we will call SC-1—is generally known (Lei et al., 2015; Gao et al., 2017), though our version is slightly more general; see Remark 5. The spectral algorithm SC-RR is proposed and analyzed in Yun and Proutiere (2014b), Gao et al. (2017) and Gao et al. (2018) for special cases of the SBM (and its degree-corrected version); the new consistency result we give for SC-RR is for the general SBM and reveals the contrast with SC-1. Previous analyses did not reveal this difference due to focusing on the special case; see Examples 1 and 2 in Section 5 for details. We also propose the new SC-RRE which has the same performance as SC-RR (assuming a proper k-means step) but is much more computationally efficient. The results that we prove for SC-RR and SC-RRE can be recast in terms of the *mean parameters* of the SBM (in contrast to SC-1), as demonstrated in Corollary 4 of Section 5.3. For an application of

this result, we refer to our work on optimal bipartite network clustering (Zhou and Amini, 2018).

1.1. Notation

Orthogonal matrices. We write $\mathbb{O}^{n \times k}$ for the set of $n \times k$ matrices with orthonormal columns. The condition $k \leq n$ is implicit in defining $\mathbb{O}^{n \times k}$. The case $\mathbb{O}^{n \times n}$ is the set of orthogonal matrices, though with some abuse of terminology we also refer to matrices in $\mathbb{O}^{n \times k}$ as orthogonal even if $k < n$. Thus, $Z \in \mathbb{O}^{n \times k}$ iff $Z^T Z = I_k$. We also note that $Z \in \mathbb{O}^{n \times k_1}$ and $U \in \mathbb{O}^{k_1 \times k}$ implies $ZU \in \mathbb{O}^{n \times k}$. The following holds:

$$\|Ux\|_2 = \|x\|_2, \quad \forall x \in \mathbb{R}^k, \ U \in \mathbb{O}^{k_1 \times k}, \tag{1}$$

for any $k_1 \geq k$. On the other hand,

$$\|U^T x\|_2 \leq \|x\|_2, \quad \forall x \in \mathbb{R}^{k_1}, \ U \in \mathbb{O}^{k_1 \times k}, \tag{2}$$

where equality holds for all $x \in \mathbb{R}^{k_1}$, iff $k_1 = k$. To see (2), let $u_1, \ldots, u_k \in \mathbb{R}^{k_1}$ be the columns of U, constituting an orthonormal sequence which can be completed to an orthonormal basis by adding say $u_{k+1}, \ldots, u_{k_1}$. Then, $\|U^T x\|_2^2 = \sum_{j=1}^{k} \langle u_j, x \rangle^2 \leq \sum_{j=1}^{k_1} \langle u_j, x \rangle^2 = \|x\|_2^2$.

Membership matrices and misclassification. We let $\mathbb{H}^{n \times k}$ denote the set of *hard* cluster labels: $\{0, 1\}$-valued $n \times k$ matrices where each row has exactly a single 1. A matrix $Z \in \mathbb{H}^{n \times k}$ is also called a membership matrix, where row i is interpreted as the membership of node i to one of k clusters (or communities). Here we implicitly assume that we have a network on nodes in $[n] = \{1, \ldots, n\}$, and there is a latent partition of $[n]$ into k clusters. In this sense, $Z_{ik} = 1$ iff node i belongs to cluster k. Given, two membership matrices $Z, Z' \in \mathbb{H}^{n \times k}$, we can consider the average misclassification rate between them, which we denote as $\overline{\mathrm{Mis}}(Z, Z')$: Letting z_i^T and $(z_i')^T$ denote the ith row of Z and Z' respectively, we have

$$\overline{\mathrm{Mis}}(Z, Z') := \min_Q \frac{1}{n} \sum_{i=1}^{n} 1\{z_i \neq Q z_i'\} \tag{3}$$

where the minimum is taken over $k \times k$ permutations matrices Q. We also let $\mathrm{Mis}_r(Z, Z')$ be the misclassification rate between the two, over the rth cluster of Z, that is, $\mathrm{Mis}_r(Z, Z') = \frac{1}{n_r} \sum_{i: \, z_i = r} 1\{z_i \neq Q^* z_i'\}$ where $n_r = \sum_{i=1}^{n} 1\{z_i = r\}$ is the size of the rth cluster of Z, and Q^* is the optimal permutation in (3). Note that in contrast to $\overline{\mathrm{Mis}}$, Mis_r is not symmetric in its two arguments. We also write $\mathrm{Mis}_\infty := \max_r \mathrm{Mis}_r$. These definitions can be extended to misclassification rates between k-means matrices introduced in Section 3.3.

2. Stochastic Block Model

We consider the general, not necessarily symmetric, Stochastic Block Model (SBM) with bi-adjacency matrix $A \in \{0, 1\}^{n_1 \times n_2}$. We assume throughout that $n_2 \geq n_1$, without loss of generality. We have membership matrices $Z_r \in \mathbb{H}^{n_r \times k_r}$ for each of the two sides $r = 1, 2$,

where $k_r \leq n_r$ denotes the number of communities on side r. Each element of A is an independent draw from a Bernoulli variable, and

$$P := \mathbb{E}[A] = Z_1 B Z_2^T, \quad B = \frac{\Psi}{\sqrt{n_1 n_2}} \tag{4}$$

where $B \in [0,1]^{k_1 \times k_2}$ is the connectivity— or the edge probability—matrix, and Ψ is a rescaled version. We also use the notation

$$A \sim \mathrm{Ber}(P) \iff A_{ij} \sim \mathrm{Ber}(P_{ij}), \text{ independent across } (i,j) \in [n_1] \times [n_2]. \tag{5}$$

Classical SBM which we refer to as *symmetric SBM* corresponds to the following modifications:

(a) A is assumed to be symmetric: Only the upper diagonal elements are drawn independently and the bottom half is filled symmetrically. For simplicity, we allow for self-loops, i.e., draw the diagonal elements from the same model. This will have a negligible effect in the arguments.

(b) $n_1 = n_2 = n$, $k_1 = k_2 = k$, $Z_1 = Z_2 = Z$.

(c) B is assumed symmetric.

We note that (4) still holds over all the elements. Directed SBM is also a special case, where (b) is assumed but not (a) or (c). That is, A is not assumed to be symmetric and all the entries are independently drawn, while B may or may not be symmetric.

We refer to P as the *mean matrix* and note that it is of rank at most $k := \min\{k_1, k_2\}$. Often $k \ll n_1, n_2$, that is P is a low-rank matrix which is the key in why spectral clustering works well for SBMs. However, the case where either $k_1 \gtrsim n_1$ or $k_2 \gtrsim n_2$ is allowed. (Here, $k_1 \gtrsim n_1$ means $k_1 \geq c n_1$ for some universal constant $c > 0$.) An extreme example of such setup can be found in Section 6.1.

We let $N_r = \mathrm{diag}(n_{r1}, \ldots, n_{rk_r})$ for $r = 1, 2$ where n_{rj} is the size of the jth cluster of Z_r; that is, N_r is a diagonal matrix whose diagonal elements are the sizes of the clusters on side r. We also consider the normalized version of N_r,

$$\bar{N}_r := N_r / n_r = \mathrm{diag}(\pi_{r1}, \ldots, \pi_{rk_r}), \tag{6}$$

collecting the cluster proportions $\pi_{rj} := n_{rj}/n_r$. Let us define

$$\bar{B} := N_1^{1/2} B N_2^{1/2} = \bar{N}_1^{1/2} \Psi \bar{N}_2^{1/2}.$$

For sparse graphs, we expect $\bar{N}_r$ and Ψ to remain stable as $n_r \to \infty$, hence $\bar{B}$ remains stable; in contrast, the entries of B itself vanish under scaling (4). See Remark 2 below for details. Throughout the paper, barred parameters refer to quantities that remain stable or slowly vary with $n_r \to \infty$. The following lemma is key in understanding spectral clustering for SBMs:

Lemma 1 *Assume that $\bar{B} = U_\psi \Sigma V_\psi^T$ is the reduced SVD of $\bar{B}$, where $U_\psi \in \mathbb{O}^{k_1 \times k}, V_\psi \in \mathbb{O}^{k_2 \times k}$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$, and $k = \min\{k_1, k_2\}$. Then,*

$$P = (\bar{Z}_1 U_\psi) \, \Sigma \, (\bar{Z}_2 V_\psi)^T \tag{7}$$

is the reduced SVD of P where $\bar{Z}_r := Z_r N_r^{-1/2}$ is itself an orthogonal matrix, $\bar{Z}_r^T \bar{Z}_r = I_{k_r}$.

Proof We first show that $Z_r^T Z_r = N_r$ which then implies that $\bar{Z}_r$ is orthogonal. Let z_{ri}^T be the ith row of Z_r and note that $Z_r^T Z_r = \sum_{i=1}^n z_{ri} z_{ri}^T$. Since $z_{ri} \in \mathbb{H}^{1 \times k_r}$, each $z_{ri} z_{ri}^T$ is a diagonal matrix with a single 1 on the diagonal at the position determined by the cluster assignment of node i on side r. Now, a little algebra on (4) shows that $P = \bar{Z}_1 \bar{B} \bar{Z}_2^T$, hence (7) holds. Since $\bar{Z}_r \in \mathbb{O}^{n_r \times k_r}$, we have $\bar{Z}_1 U_\psi \in \mathbb{O}^{n_1 \times k}$ and $\bar{Z}_2 V_\psi \in \mathbb{O}^{n_2 \times k}$ showing that (7) is in fact a (reduced) SVD of P. ∎

When dealing with the symmetric SBM, we will drop the subscript r from all the relevant quantities; for example, we write $N = N_1 = N_2$, $\bar{Z} = \bar{Z}_1 = \bar{Z}_2$, $\pi_j = \pi_{1j} = \pi_{2j}$, and so on.

Remark 1 (Reduced versus truncated) *The term* reduced SVD *in Lemma 1 (also known as compact SVD) means that we reduce the orthogonal matrices in a full SVD by removing the columns corresponding to zero singular values. The number of columns of the resulting matrices will be equal to the rank of the underlying matrix (i.e., both $\bar{Z}_1 U_\psi$ and $\bar{Z}_2 V_\psi$ will have $k = \min\{k_1, k_2\}$ columns in the case of P). Hence, a reduced SVD is still an exact SVD. Later, we will use the term* truncated SVD *to refer to an "approximation" of the original matrix by a lower rank matrix obtained by further removing columns corresponding to small nonzero singular values (starting from a reduced SVD). Hence, a truncated SVD is only an approximation of the original matrix.*

Remark 2 (Scaling and sparsity) *As can be seen from the above discussion, the normalization in (4) is natural for studying spectral clustering. In the symmetric case, where $n_1 = n_2 = n$, the normalization reduces to $B = \Psi/n$, which is often assumed when studying sparse SBMs by requiring that either $\|\Psi\|_\infty$ is $O(1)$ or grows slowly with n. To see why this implies a sparse network, note that the expected average degree of the symmetric SBM (under this scaling) is*

$$\frac{1}{n} 1^T P 1 = \frac{1}{n} 1^T N B N 1 = 1^T \bar{N} \Psi \bar{N} 1 = \sum_{i,j=1}^k \pi_i \pi_j \Psi_{ij} =: d_{av}$$

using $1_n^T Z = 1_k^T N$. (Here and elsewhere, 1 is the vector of all ones of an appropriate dimension; we write 1_n if we want to emphasize the dimension n.) Thus, the growth of the average expected degree, d_{av}, is the same as Ψ, and as long as Ψ is $O(1)$ or grows very slowly with n, the network is sparse. Alternatively, we can view the expected density of the network (the expected number of edges divided by the total number of possible edges) as a measure of sparsity. For the symmetric case, the expected density is $(\frac{1}{2} 1^T P 1)/\binom{n}{2} \sim d_{av}/n$ and is $O(n^{-1})$ if $d_{av} = O(1)$. Similar observations hold in the general bipartite case if we

let $n = \sqrt{n_1 n_2}$, the geometric mean of the dimensions. The expected density of the bipartite network under the scaling of (4) is

$$\frac{1^T P 1}{n^2} = \frac{1}{n^2} 1^T N_1 B N_2 1 = \frac{1}{n} 1^T \bar{N}_1 \Psi \bar{N}_2 1 = \frac{d_{av}}{n}, \quad (n = \sqrt{n_1 n_2})$$

where $d_{av} := 1^T \bar{N}_1 \Psi \bar{N}_2 1 = \sum_{i,j} \pi_{1i} \pi_{2j} \Psi_{ij}$ can be thought of as the analog of the expected average degree in the bipartite case. As long as $\|\Psi\|_\infty$ grows slowly relative to $n = \sqrt{n_1 n_2}$, the bipartite network is sparse.

3. Analysis steps

Throughout, we focus on recovering the row clusters. Everything that we discuss goes through, with obvious modifications, for recovering the column clusters. Recalling the decomposition (7), the idea of spectral clustering in the context of SBMs is that $\bar{Z}_1 U_\psi$ has enough information for recovering the clusters and can be obtained by computing a reduced SVD of P. In particular, applying a k-means type clustering on the rows of $\bar{Z}_1 U_\psi$ should recover the cluster labels. On the other hand, the actual random adjacency matrix, A, is concentrated around the mean matrix P, after proper regularization if need be. We denote this potentially *regularized version as* A_{re}. Then, by the spectral perturbation theory, if we compute a reduced SVD of $A_{\mathrm{re}} = \hat{Z}_1 \hat{\Sigma} \hat{Z}_2^T$ where $\hat{Z}_r \in \mathbb{O}^{n_r \times k}$, $r = 1, 2$ and $\hat{\Sigma}$ is diagonal, we can conclude that $\hat{Z}_1$ concentrates around $\bar{Z}_1 U_\psi$. Hence, applying a continuous k-means algorithm on $\hat{Z}_1$ should be able to recover the labels with a small error.

3.1. Analysis sketch

Let us sketch the argument above in more details. A typical approach in proving consistency of spectral clustering consists of the following steps:

1. We replace A with a properly regularized version A_{re}. We provide the details for one such regularization in Theorem 2 (Section 4). However, the only property we require of the regularized version is that it concentrates, with high probability, around the mean of A, at the following rate (assuming $n_2 \geq n_1$):

$$\|A_{\mathrm{re}} - \mathbb{E}[A]\|_{\mathrm{op}} \leq C\sqrt{d}, \quad \text{where} \quad d = \sqrt{\frac{n_2}{n_1}} \|\Psi\|_\infty. \tag{8}$$

Here and throughout $\|\cdot\|_{\mathrm{op}}$ is the $\ell_2 \to \ell_2$ operator norm and $\|\Psi\|_\infty = \max_{ij} \Psi_{ij}$.

2. We pass from A_{re} and $P = \mathbb{E}[A]$ to their (symmetrically) dilated versions $A_{\mathrm{re}}^\dagger$ and $P^\dagger$. The symmetric dilation operator will be given in (13) (Section 3.2) and allows us to use spectral perturbation bounds for symmetric matrices. A typical final result of this step is

$$\|\hat{Z}_1 - \bar{Z}_1 U_\psi Q\|_F \leq \frac{C_2}{\sigma_k} \sqrt{kd}, \quad \text{w.h.p.} \tag{9}$$

for some $Q \in \mathbb{O}^{k \times k}$. We recall that $\|\cdot\|_F$ is the Frobenius norm. Here, σ_k is the smallest nonzero singular value of $\bar{B}$ as defined in Lemma 1. The form of (9) will be different

if instead of $\hat{Z}_1$ one considers other objects as the end result of this step; see Section 5 (e.g., (34)) for instances of such variations. The appearance of Q is inevitable and is a consequence of the necessity of *properly aligning* the bases of spectral subspaces, before they can be compared in Frobenius norm (cf. Lemma 3). Nevertheless, the growing stack of orthogonal matrices on the RHS of $\bar{Z}_1$ has little effect on the performance of row-wise k-means, as we discuss shortly.

3. The final step is to analyze the effect of applying a k-means algorithm to $\hat{Z}_1$. Here, we introduce the concept of a *k-means matrix*, one whose rows take at most k distinct values. (See Section 3.3 for details). A k-means algorithm $\mathscr{K}$ takes a matrix $\hat{X} \in \mathbb{R}^{n \times m}$ and outputs a k-means matrix $\mathscr{K}(\hat{X}) \in \mathbb{R}^{n \times m}$. Our focus will be on k-means algorithms with the following property: If $X^* \in \mathbb{R}^{n \times m}$ is a k-means matrix, then for some constant $c > 0$,

$$\|\hat{X} - X^*\|_F^2 \le \varepsilon^2 \implies \overline{\mathrm{Mis}}(\mathscr{K}(\hat{X}), X^*) \le c\,\varepsilon^2/(n\delta^2). \tag{10}$$

where $\delta^2 = \delta^2(X^*)$ is the minimum center separation of X^* (cf. Definition 2), and $\overline{\mathrm{Mis}}$ is the average misclassification rate between two k-means matrices. For future reference, we refer to property (10) as the *local quadratic continuity (LQC)* of algorithm $\mathscr{K}$; see Remark 6 for the rationale behind the naming. As will become clear in Section 3.3, k-means matrices encode both the cluster label information and cluster center information, and these two pieces can be recovered from them in a lossless fashion. Thus, it makes sense to talk about misclassification rate between k-means matrices, by interpreting it as a statement about their underlying label information. In Section 3.3, we will discuss k-means algorithms that satisfy (10).

The preceding three steps of the analysis follow the three steps of a general spectral clustering algorithm, which we refer to as *regularization*, *spectral truncation* and *k-means* steps, respectively. Recalling the definition of cluster proportions, let us assume for some $\beta_r \ge 1$,

$$\max_{(t,s):\, t\neq s} \frac{2}{\pi_{rt}^{-1} + \pi_{rs}^{-1}} \;\le\; \frac{\beta_r}{k_r}, \quad r = 1, 2. \tag{A1}$$

The LHS is the maximum *harmonic mean* of pairs of distinct cluster proportions. For balanced clusters, we have $\pi_{rt} = 1/k_r$ for all $t \in [k_r]$ and we can take $\beta_r = 1$. In general, β_r measures the deviation of the clusters (on side r) from balancedness. The following is a prototypical consistency theorem for a spectral clustering algorithm:

Theorem 1 (Prototype SC consistency) *Consider a spectral algorithm with a k-means step satisfying* (10), *and the "usual" spectral truncation step, applied to a regularized bi-adjacency matrix A_{re} satisfying concentration bound* (8). *Let $\mathscr{K}(\hat{Z}_1)$ be the resulting estimate for membership matrix Z_1, and assume $k_1 = k =: \min\{k_1, k_2\}$. Then, under the SBM model of Section 2 and assuming* (A1), *w.h.p.,*

$$\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1), \bar{Z}_1) \;\lesssim\; \beta_1\Big(\frac{d}{\sigma_k^2}\Big).$$

Here, and in the sequel, "with high probability", abbreviated w.h.p., means with probability at least $1 - n^{-c_1}$ for some universal constant $c_1 > 0$. The notation $f \lesssim g$ means $f \leq c_2 g$ where $c_2 > 0$ is a universal constant. In addition, $f \asymp g$ means $f \lesssim g$ and $g \lesssim f$.

Proof By assumption, concentration bound (8) holds. By Lemma 3 in Section 3.2, (8) implies (9) for the usual truncation step. Let $O := U_\psi Q$ and $\varepsilon^2 = C_2^2 \, kd/\sigma_k^2$ so that (9) reads $\|\hat{Z}_1 - \bar{Z}_1 O\|_F \leq \varepsilon^2$. The k-means step satisfies (10) by assumption. Applying (10) with $\hat{X} = \hat{Z}_1$, $X^* = \bar{Z}_1 O$ and ε^2 defined earlier leads to

$$\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1), \bar{Z}_1) = \overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1), \bar{Z}_1 O) \lesssim \frac{\varepsilon^2}{n_1 \delta^2}. \tag{11}$$

It remains to calculate the minimum center separation of $X^* = \bar{Z}_1 O \in \mathbb{R}^{n_1 \times k}$, where $\bar{Z}_1 \in \mathbb{O}^{n_1 \times k_1}$ and $O := U_\psi Q \in \mathbb{O}^{k_1 \times k}$. We have

$$\delta^2 = \delta^2(\bar{Z}_1 O) = \delta^2(\bar{Z}_1) = \min_{t \neq s} \|n_{1t}^{-1/2} e_t - n_{1s}^{-1/2} e_s\|_2^2 = \min_{t \neq s} (n_{1t}^{-1} + n_{1s}^{-1})$$

where $e_s \in \mathbb{R}^{k_1}$ is the sth standard basis vector. The second equality uses invariance of δ^2 to right-multiplication by a square orthogonal matrix. This is a consequence of $\|u^T O - v^T O\|_2 = \|u - v\|_2$ for $u, v \in \mathbb{R}^{k_1}$ and $O \in \mathbb{O}^{k_1 \times k}$ when $k_1 = k$; see (1). The third equality is from the definition $\bar{Z}_1 = Z_1 N_1^{-1/2}$. Using (A1),

$$(n_1 \delta^2)^{-1} \leq \max_{t \neq s} (\pi_{1t}^{-1} + \pi_{1s}^{-1})^{-1} \leq \frac{\beta_1}{2k_1}.$$

Thus, $\frac{\varepsilon^2}{n_1 \delta^2} \lesssim \beta_1 \frac{k}{k_1} \frac{d}{\sigma_k^2}$ which gives the result in light of (11) and assumption $k_1 = k$. ∎

The case $k_1 > k$ will be discussed in Section 6.2. For (10) to hold for a k-means algorithm, one usually requires some additional constraints on $\varepsilon^2/(n\delta^2)$, ensuring that this quantity is small. We will restate Theorem 1 with such conditions explicitly once we consider the details of some k-means algorithms. For now, Theorem 1 should be thought of as a general blueprint, with specific variations obtained in Section 5 for various spectral clustering algorithms.

Remark 3 *To see that Theorem 1 is a consistency result, consider the typical case where $\beta_1 \asymp 1$, and $\sigma_k \asymp d$, so that $\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1), \bar{Z}_1) = O(d^{-1})$. Then, as long as $d \to \infty$, i.e., the average degree of the network grows with n, assuming $n_1 \asymp n_2 \asymp n$ (for some n), we have $\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1), \bar{Z}_1) = o(1)$, i.e., the average misclassification rate vanishes with high probability. One might ask why $\sigma_k \asymp d$ is reasonable. Consider the typical case where $\Psi = \rho_n \Psi'$ for some constant matrix Ψ' and a scalar parameter ρ_n that captures the dependence on n. This setup is common in network modeling (Bickel and Chen, 2009). Then, $d \asymp \|\Psi\|_\infty = \rho_n \|\Psi'\|_\infty$ and $\sigma_k = \rho_n \sigma_k'$ where σ_k' is defined based on Ψ' and hence constant. It follows that $d \asymp \rho_n \asymp \sigma_k$ and $\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1), \bar{Z}_1) = o(\rho_n^{-1})$. Note that, in general, ρ_n can grow as fast as n (cf. (4)). More specific examples are given in Section 5.*

Remark 4 *We note that bounds similar to (9) appear in the recent literature (Abbe et al., 2017; Cape et al., 2018) but for the $2 \to \infty$ operator norm, as opposed to the Frobenius norm,*

assuming that the network is dense enough. These bounds can be used to show that spectral clustering achieves exact recovery when the degree grows as $\log n$ for the two-community case (Abbe et al., 2017). A similar exact recovery result for general K is obtained in (Su et al., 2017) assuming a degree growth of $(\log n)^a$ for sufficiently large a. Whether the $2 \to \infty$ bounds can be used to sharpen the misclassification rate of spectral clustering below the $\log n$ regime is open.

Remark 5 *Condition (A1) is more relaxed that what is commonly assumed in the literature (though the proof is the same). Stating the condition as a harmonic mean allows one to have similar results as the balanced case when one cluster is large, while others remain more or less balanced. For example, let $\pi_{r1} = 1 - c$ for some constant $c \in (0,1)$, say $c = 0.4$, and let $\pi_{rt} = c/(k_r - 1)$ for $t \neq 1$. Then, we have for $s \neq t$*

$$\frac{2}{\pi_{rt}^{-1} + \pi_{rs}^{-1}} \leq 2\min\{\pi_{rt}, \pi_{rs}\} = \frac{2c}{k_r - 1} \leq \frac{4}{k_r}$$

assuming $k_r \geq 2$. Hence (A1) holds with $\beta_r = 4$. Note that as k_r is increased, all but one cluster get smaller.

Remark 6 (LQC naming) *The rationale behind the naming of (10) is as follows: Let $\Delta(X,Y) := \frac{1}{\sqrt{n}}\|X - Y\|_F$ be the metric induced by the normalized Frobenius norm on the space of $n \times m$ matrices. Assume that X^* is a k-means matrix and that k-means matrices are fixed points of algorithm $\mathscr{K}$, hence $X^* = \mathscr{K}(X^*)$. Then, by taking the infimum over ε^2, (10) can be written as*

$$\overline{\mathrm{Mis}}\left(\mathscr{K}(\hat{X}), \mathscr{K}(X^*)\right) \leq \omega\left(\Delta(\hat{X}, X^*)\right) \tag{12}$$

where $\omega(t) = ct^2/\delta^2(X^*)$, showing that $\mathscr{K}$ is continuous w.r.t. the two (pseudo)-metrics, locally at X^*, with a quadratic modulus of continuity. Note that this continuity is only required to hold around a k-means matrix X^* (and not in general). Another reason for the "locality" is that such statements often only hold for sufficiently small $\Delta(\hat{X}, X^*)$.

In this rest of this section, we fill in some details of the last two steps of the plan sketched above, deferring Step 1 to Section 4.

3.2. SV truncation (Step 2)

We now show how the concentration bound (8) implies the deviation bound (9) for the usual spectral truncation step. Consider the *symmetric dilation* operator : $\mathbb{R}^{n_1 \times n_2} \to \mathbb{S}^{n_1+n_2}$, given by

$$P^{\dagger} := \begin{pmatrix} 0 & P \\ P^T & 0 \end{pmatrix}, \tag{13}$$

where $\mathbb{S}^n$ is the set of symmetric $n \times n$ matrices. This operator will be very useful in translating the results between the symmetric and non-symmetric cases. Let us collect some of its properties:

Lemma 2 (Symmetric Dilation) *Let $P \in \mathbb{R}^{n_1 \times n_2}$ have a reduced SVD given by $P = U\Sigma V^T$ where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ is a $k \times k$ nonnegative diagonal matrix . Then,*

(a) $P^\dagger$ has a reduced EVD given by

$$P^\dagger = W \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix} W^T, \quad W = \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ V & -V \end{pmatrix} \in \mathbb{O}^{(n_1+n_2) \times 2k}.$$

(b) $P \mapsto P^\dagger$ is a linear operator with $\|P^\dagger\| = \|P\|$ and $\|P^\dagger\|_F = \sqrt{2}\|P\|_F$.

(c) The gap between k top (signed) eigenvalues of $P^\dagger$ and the rest of its spectrum is $2\sigma_k$.

Proof Part (a) can be verified directly (e.g. $W^T W = I_{2k}$ follows from $U^T U = V^T V = I_k$) and part (c) follows by noting that $\sigma_j \geq 0$ for all j. Part (b) also follows directly from part (a), using unitary-invariance of the two norms. ∎

In addition, let us define a singular value (SV) truncation operator $\mathscr{T}_k : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ that takes a matrix A with SVD $A = \sum_i \sigma_i u_i v_i^T$ to the matrix

$$A^{(k)} := \mathscr{T}_k(A) := \sum_{i=1}^{k} \sigma_i u_i v_i^T. \tag{14}$$

In other words, $\mathscr{T}_k$ keeps the largest k singular values (and the corresponding singular vectors) and zeros out the rest. Recall that we order singular values in nonincreasing fashion $\sigma_1 \geq \sigma_2 \geq \cdots$. We also refer to (14) as the *k-truncated SVD* of A. Using the dilation and the Davis–Kahan (DK) theorem for symmetric matrices (Bhatia, 2013, Theorem VII.3.1), we have:

Lemma 3 *Let $A_{re}^{(k)} = \hat{Z}_1 \hat{\Sigma} \hat{Z}_2^T$ be the k-truncated SVD of A_{re} and assume that the concentration bound (8) holds. Let $\bar{Z}_1 U_\psi$ be given by the reduced SVD of P in (7). Then, the deviation bound (9) holds for some $k \times k$ orthogonal matrix Q, and $C_2 = 2C$.*

Later, in Section 5, we introduce an alternative spectral truncation scheme. It is worth comparing the above lemma to Lemma 6 which establishes a similar result for the alternative truncation.

Remark 7 (Symmetric case) *When P is symmetric one can still use the dilation operator. In this case, since P itself is symmetric, it has an eigenvalue decomposition (EVD), say $P = U\Lambda U^T$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ is the diagonal matrix of the eigenvalues of P. Since these eigenvalues could be negative, there is a slight modification needed to go from the EVD to the SVD of P. Let s_i be the sign of λ_i and set $S = \mathrm{diag}(s_i, i = 1, \ldots, k)$. Then, it is not hard to see that with $V = US$ and $\Sigma = \Lambda S = \mathrm{diag}(|\lambda_i|, i = 1, \ldots, k)$, we obtain the SVD $P = U\Sigma V^T$. In other words, all the discussion in this section, and in particular Lemma 3 hold with $V = US$ and $\sigma_i = |\lambda_i|$. The special case of Lemmas 2 and 3 for the symmetric case appears in Lei et al. (2015). These observations combined with the fact that the concentration inequality discussed in Section 4 holds in the symmetric case leads to the following conclusion: All the results discussed in this paper apply to the symmetric SBM,*

for the version of the adjacency-based spectral clustering that sorts the eigenvalues based on their absolute values. This is the most common version of spectral algorithms in use. On the other hand, one gets a different behavior for the algorithm that considers the top k (signed) eigenvalues. We have borrowed the term "symmetric dilation" from Tropp (2015) where these ideas have been successfully used in translating matrix concentration inequalities to the symmetric case.

3.3. k-means algorithms (Step 3)

Let us now give the details of the third and final step of the analysis. We introduce some notations and concepts that help in the discussion of k-means (type) algorithms.

k-means matrices. Recall that $\mathbb{H}^{n \times k}$ denotes the set of *hard* (cluster) labels: $\{0, 1\}$-valued $n \times k$ matrices where each row has exactly a single 1. Take $Z \in \mathbb{H}^{n \times k}$. A related notion is that of a *cluster* matrix $Y = ZZ^T \in \{0, 1\}^n$ where each entry denotes whether the corresponding pair are in the same cluster. Relative to Z, Y loses the information about the ordering of the cluster labels. We define the class of *k-means matrices* as follows:

$$
\begin{aligned}
\mathbb{M}^k_{n,m} &:= \{X \in \mathbb{R}^{n \times m} : X \text{ has at most } k \text{ distinct rows}\} \\
&= \{ZR : Z \in \mathbb{H}^{n \times k}, R \in \mathbb{R}^{k \times m}\}.
\end{aligned}
\tag{15}
$$

The rows of R, which we denote as r_i^T, play the role of cluster centers. Let us also denote the rows of X as x_i^T. The second equality in (15) is due to the following correspondence: Any matrix $X \in \mathbb{M}^k_{n,m}$ uniquely identifies a *cluster* matrix $Y \in \{0, 1\}^{n \times n}$ via, $Y_{ij} = 1$ iff $x_i = x_j$. This in turn "uniquely" identifies a label matrix Z up to $k!$ permutation of the labels. From Z, we "uniquely" recover R, with the convention of setting rows of R for which there is no label equal to zero. (This could happen if X has fewer than k distinct rows.)

With these conventions, there is a one-to-one correspondence between $X \in \mathbb{M}^k_{n,m}$ and $(Z, R) \in \mathbb{H}^{n \times k} \times \mathbb{R}^{k \times m}$, up to label permutations. That is, (Z, R) and (ZQ, QR) are considered equivalent for any permutation matrix Q. The correspondence allows us to talk about a (relative) misclassification rate between two k-means matrices: If $X_1, X_2 \in \mathbb{M}^k_{n,m}$ with membership matrices $Z_1, Z_2 \in \mathbb{H}^{n \times k}$, respectively, we set

$$
\overline{\text{Mis}}(X_1, X_2) := \overline{\text{Mis}}(Z_1, Z_2).
\tag{16}
$$

k-means as projection. Now consider a general $\hat{X} \in \mathbb{R}^{n \times m}$. The classical k-means problem can be thought of as projecting $\hat{X}$ onto $\mathbb{M}^k_{n,m}$, in the sense of finding a nearest member of $\mathbb{M}^k_{n,m}$ to $\hat{X}$ in Frobenius norm. Let us write $d_F(\cdot, \cdot)$ for the distance induced by the Frobenius norm, i.e., $d_F(\hat{X}, X) = \|\hat{X} - X\|_F$. The k-means problem is that of solving the following optimization:

$$
d_F(\hat{X}, \mathbb{M}^k_{n,m}) := \min_{X \in \mathbb{M}^k_{n,m}} d_F(\hat{X}, X).
\tag{17}
$$

The arguments to follow go through for any distance on matrices that has a ℓ_2 decomposition over the rows:

$$
d_F(\hat{X}, X)^2 = \sum_{i=1}^n d(\hat{x}_i, x_i)^2,
\tag{18}
$$

where x_i^T and $\hat{x}_i^T$ are the rows of X and $\hat{X}$ respectively, and $d(\hat{x}_i, x_i)$ is some distance over vectors in $\mathbb{R}^d$. For the case of the Frobenius norm: $d(\hat{x}_i, x_i) = \|\hat{x} - x_i\|_2$, the usual ℓ_2 distance. This is the primary case we are interested in, though the result should be understood for the general case of (18). Since solving the k-means problem (17) is NP-hard, one can look for approximate solutions:

Definition 1 *A κ-approximate k-means solution for $\hat{X}$ is a matrix $\widetilde{X} \in \mathbb{M}_{n,m}^k$ that achieves κ times the optimal distance:*

$$d_F(\hat{X}, \widetilde{X}) \leq \kappa \, d_F(\hat{X}, \mathbb{M}_{n,m}^k). \tag{19}$$

We write $\mathcal{P}_\kappa : \mathbb{R}^{n \times m} \mapsto \mathbb{M}_{n,m}^k$ for the (set-valued) function that maps matrices $\hat{X}$ to κ-approximate solutions $\widetilde{X}$.

An equivalent restatement of (19) is

$$d_F(\hat{X}, \widetilde{X}) \leq \kappa \, d_F(\hat{X}, X), \quad \forall X \in \mathbb{M}_{n,m}^k. \tag{20}$$

Note that $\mathcal{P}_\kappa(\hat{X}) = \{\widetilde{X} \in \mathbb{M}_{n,m}^k : \ \widetilde{X} \text{ satisfies (20)}\}$. Our goal is to show that whenever $\hat{X}$ is close to some $X^* \in \mathbb{M}_{n,m}^k$, then any κ-approximate k-means solution based on it, namely $\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})$ will be close to X^* as well. This is done in two steps:

1. If the distance $d_F(X, \widetilde{X})$ between two k-means matrices $X, \widetilde{X} \in \mathbb{M}_{n,m}^k$ is small, then their relative misclassification rate $\overline{\mathrm{Mis}}(X, \widetilde{X})$ is so.

2. If a general matrix $\hat{X} \in \mathbb{R}^{n \times d}$ is close to a k-means matrix $X \in \mathbb{M}_{n,m}^k$, then so is its κ-approximate k-means projection. More specifically,

$$d_F(\widetilde{X}, X) \leq (1 + \kappa) \, d_F(\hat{X}, X), \quad \forall \widetilde{X} \in \mathcal{P}_\kappa(\hat{X}), \tag{21}$$

which follows from the triangle inequality $d_F(\widetilde{X}, X) \leq d_F(\widetilde{X}, \hat{X}) + d_F(\hat{X}, X)$ and (20).

Combining the two steps (taking $X = X^*$), we will have the result.

Let us now give the details of the first step above. For this result, we need the key notion of center separation. k-means matrices have more information that just a membership assignment. They also contain an encoding of the relative positions of the clusters, and hence the minimal pairwise distance between them, which is key in establishing a misclassification rate.

Definition 2 (Center separation) *For any $X \in \mathbb{M}_{n,m}^k$, let us denote its centers, i.e. distinct rows, as $\{q_r(X), r \in [k]\}$, and let*

$$\delta_r(X) = \min_{\ell: \, \ell \neq r} d(q_\ell(X), q_r(X)), \quad \delta_\wedge(X) = \min_r \delta_r(X). \tag{22}$$

In addition, let $n_r(X)$ be the number of nodes in cluster r according to X, and $n_\wedge(X) = \min_r n_r(X)$, the minimum cluster size.

If X has $m < k$, the convention would be to let $q_k(X) = 0$ for $k = m+1, \ldots, k$. We usually do not work with these degenerate cases. Implicit in the above definition is an enumeration of the clusters of X. We note that definition of $\delta_r = \delta_r(X)$ in (22) implies

$$d(q_\ell(X), q_r(X)) \geq \max\{\delta_\ell, \delta_r\}, \quad \forall (r, \ell) : r \neq \ell. \tag{23}$$

We are now ready to show that that any algorithm that computes a κ-approximate solution to the k-means problem (17) (where κ is some constant) satisfies the LQC property (10) needed in the last step of the analysis sketched in Section 3.1. Recall that $\mathrm{Mis}_r(X; \widetilde{X})$ is the misclassification rate over the rth cluster of X (Section 1.1).

Proposition 1 *Let $X, \widetilde{X} \in \mathbb{M}_{n,m}^k$ be two k-means matrices, and write $n_r = n_r(X)$, $n_\wedge = n_\wedge(X)$ and $\delta_r = \delta_r(X)$. Assume that $d_F(X, \widetilde{X}) \leq \varepsilon$ and*

(a) *X has exactly k nonempty clusters, and*

(b) *$c_r^{-2} \varepsilon^2 / (\delta_r^2 n_r) < 1$ for $r \in [k]$, and constants $c_r > 0$ such that $c_r + c_\ell \leq 1$, $r \neq \ell$.*

Then, $\widetilde{X}$ has exactly k clusters and

$$\mathrm{Mis}_r(X; \widetilde{X}) \leq \frac{c_r^{-2} \varepsilon^2}{n_r \, \delta_r^2}, \quad \forall r \in [k]. \tag{24}$$

In particular, under the conditions of Proposition 1 with $c_r = 1/2$, we have

$$\mathrm{Mis}_\infty(X, \widetilde{X}) \leq \frac{4\,\varepsilon^2}{\min_r n_r \delta_r^2} \leq \frac{4\,\varepsilon^2}{n_\wedge \delta_\wedge^2}, \qquad \overline{\mathrm{Mis}}(X, \widetilde{X}) \leq \frac{4\,\varepsilon^2}{n\, \delta_\wedge^2}.$$

where the second one follows from the identity $\overline{\mathrm{Mis}}(X, \widetilde{X}) = \sum_{r=1}^k (n_r/n) \, \mathrm{Mis}_r(X, \widetilde{X})$. Combining Proposition 1 with (21), we obtain the following corollary:

Corollary 1 *Let $X^* \in \mathbb{M}_{n,m}^k$ be a k-means matrix, and write $n_r = n_r(X^*)$, $n_\wedge = n_\wedge(X^*)$ and $\delta_r = \delta_r(X^*)$. Assume that $\hat{X} \in \mathbb{R}^{n \times m}$ is such that $d_F(X^*, \hat{X}) \leq \varepsilon$ and*

(a) *X^* has exactly k nonempty clusters, and*

(b) *$c_r^{-2} (1+\kappa)^2 \varepsilon^2 / (\delta_r^2 n_r) < 1$ for $r \in [k]$, and constants $c_r > 0$ such that $c_r + c_\ell \leq 1$, $r \neq \ell$.*

Then, any $\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})$ has exactly k clusters and

$$\mathrm{Mis}_r(X^*; \mathcal{P}_\kappa(\hat{X})) \leq \frac{c_r^{-2} (1 + \kappa)^2 \varepsilon^2}{n_r \, \delta_r^2}, \quad \forall r \in [k]. \tag{25}$$

As before, $\mathrm{Mis}_r(X^*, \mathcal{P}_\kappa(\hat{X}))$ should be interpreted as $\max_{\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})} \mathrm{Mis}_r(X^*, \widetilde{X})$, that is, the result hold for any κ-approximate k-means solution for $\hat{X}$. In particular, under the conditions of Corollary 1 with $c_r = 1/2$, we have

$$\overline{\mathrm{Mis}}(X^*, \mathcal{P}_\kappa(\hat{X})) \leq 4(1 + \kappa)^2 \frac{\varepsilon^2}{n\, \delta_\wedge^2}. \tag{26}$$

A similar bound holds for Mis_∞. Note that (26) is of the desired form needed in (10).

Remark 8 *Corollary 1 shows that any constant-factor approximation to the k-means problem (17) satisfies the LCQ property (10). Such approximations can be computed in polynomial time; see for example Kumar et al. (2004). One can also use Lloyd's algorithm with* `kmeans++` *initialization to get a $\kappa \lesssim \log k$ approximation (Arthur and Vassilvitskii, 2007). Both Kumar et al. (2004) and Arthur and Vassilvitskii (2007) give constant probability approximations, hence if such algorithms are used in our subsequent results, "w.h.p." should be interpreted as with high constant probability (rather than $1 - o(1)$). Recently Lu and Zhou (2016) has shown that Lloyd's algorithm with random initialization can achieve near optimal misclassification rate in certain random models. Their analysis is complementary to ours in that we establish and use the LQC property (10) which uniformly holds for any input matrix $\hat{X}$ and we do not consider specific algorithms. The core ideas of our analysis in this section are borrowed from Lei et al. (2015) and Jin (2015). It is worth noting that there are other algorithms that turn a general matrix into a k-means matrix, without trying to approximate the k-means problem (17), and still satisfy the LQC. We will give one such example, Algorithm 6, in Appendix B. Interestingly, in contrast to the k-means approximation algorithms, the LQC guarantee for Algorithm 6 is not probabilistic.*

4. Regularization and concentration

Here, we provide the details of Step 1, namely, the concentration of the regularized adjacency matrix. We start by a slight generalization of the results of Le et al. (2017) to the rectangular case:

Theorem 2 *Assume $n_1 \leq n_2$ and let $A \in \{0,1\}^{n_1 \times n_2}$ have independent Bernoulli entries with mean $\mathbb{E}[A_{ij}] = p_{ij}$. Take $d \geq \max_{ij} n_2 \, p_{ij}$. Pick any subsets $\mathcal{I}_1 \subset [n_1]$ and $\mathcal{I}_2 \subset [n_2]$ such that*

$$|\mathcal{I}_1| \leq 10n_2/d, \quad and \quad |\mathcal{I}_2| \leq 10n_2/d \tag{27}$$

and fix some $d' > 0$. Define a regularized adjacency matrix A_{re} as follows:

(a) *Set $[A_{re}]_{ij} = A_{ij}$ for all $(i,j) \in \mathcal{I}_1^c \times \mathcal{I}_2^c$.*

(b) *Set $[A_{re}]_{ij}$ arbitrarily when $i \in \mathcal{I}_1$ or $j \in \mathcal{I}_2$, but subject to the constraints $\|[A_{re}]_{i*}\|_1 \leq d'$ and $\|[A_{re}]_{*j}\|_1 \leq d'$ for all $i \in \mathcal{I}_1$ and $j \in \mathcal{I}_2$.*

Then, for any $r \geq 1$, with probability at least $1 - n_2^{-r}$, the new adjacency matrix A_{re} satisfies

$$\|A_{re} - \mathbb{E}[A]\|_{op} \leq C_2 \, r^{3/2} (\sqrt{d} + \sqrt{d'}). \tag{28}$$

The same result holds if in step (b) one uses ℓ_2 norm instead of ℓ_1.

Theorem 2 follows directly from the non-symmetric (i.e., directed) version of (Le et al., 2017, Theorem 2.1), by padding A with rows of zeros to get a square $n_2 \times n_2$ matrix. The result then follows by the same argument as in (Le et al., 2017, Theorem 2.1). The term "arbitrary" in the statement of the theorem includes any reduction even if the scheme is stochastic and depends on A itself. This feature will be key in developing data-driven schemes.

To apply Theorem 2, take $\mathcal{I}_1 = \{i \in [n_1] : \sum_j A_{ij} > 2d\}$ and $\mathcal{I}_2 = \{j \in [n_2] : \sum_i A_{ij} > 2d\}$, i.e., the set of rows and columns with degrees larger than $2d$. It is not hard to see that if d is any upper bound on the expected row and column degrees, then the sizes of these sets satisfy (27) with high probability. This follows, for example, from the same argument as that leading to (64) in Appendix A.2. Recalling the scaling of the connectivity matrix in (4), taking $d = n_2\|P\|_\infty$ gives an upper bound on the expected row and column degrees, assuming $n_2 \geq n_1$. Thus, we can apply Theorem 2 with $d' = d = \sqrt{n_2/n_1}\|\Psi\|_\infty = n_2\|P\|_\infty$, to obtain the desired concentration bound (8). Note that this concentration result does not require $d = n_2\|P\|_\infty$ to satisfy any lower bound (such as $d = \Omega(\log n_2)$). See Remarks 9 and 10 for the significance of this fact.

A disadvantage of the regularization scheme described in Theorem 2 is the required knowledge of a good upper bound on $d = n_2\|P\|_\infty$. In the next section, under a SBM with some mild regularity assumptions, we develop a fully data-driven scheme with the same guarantees as those of Theorem 2.

4.1. Data-driven regularization

We now describe a regularization scheme that is data-driven and does not require the knowledge of d as in Theorem 2. Let us write $D_i = \sum_{j=1}^{n_2} A_{ij}$ for the degree of node i on side 1 and $\overline{D} = \frac{1}{n_1}\sum_{i=1}^{n_1} D_i$ for the average degree on side 1. Consider the following order statistics for the degree sequence:

$$D_{(1)} \geq D_{(2)} \geq \cdots \geq D_{(n_1)}. \tag{29}$$

The idea is that under a block model, we can achieve concentration (8) by reducing the row degrees that are roughly above $D_{(\alpha)}$ for $\alpha = \lfloor n_1/\overline{D} \rfloor$ (and similarly for the columns). The overall scheme is described in Algorithm 1. The algorithm has a tuning parameter τ; however, we will show that taking $\tau = 3$ is enough. This constant is for convenience and has no special meaning. Note that the regularization described in Step 6 is a special case of that in Step 5; it is equivalent to first reducing the row degrees by setting $[A_{\mathrm{re}}]_{i*} \leftarrow A_{i*}\widehat{d}_1/D_i$ for $i \in \widehat{\mathcal{I}}_1$ followed by reducing the column degrees $[A_{\mathrm{re}}]_{*j} \leftarrow [A_{\mathrm{re}}]_{*j}\widehat{d}_2/D_j$ for $j \in \widehat{\mathcal{I}}_2$, or vice versa. Alternatively, we can replace the ℓ_1 norm constraints in Step 5 with ℓ_2 norm version and replace Step 6 with $[A_{\mathrm{re}}]_{ij} = A_{ij}\sqrt{w_i w_j'}$ for all $i \in [n_1], j \in [n_2]$.

Let $p_{ij} = \mathbb{E}[A_{ij}]$ and define $d_i := \mathbb{E}[D_i] = \sum_j p_{ij}$ and

$$\bar{d} = \frac{1}{n_1}\sum_{i=1}^{n} d_i, \quad d_{\max} := \max_i d_i, \quad d := n_2 \cdot \max_{i,j} p_{ij}. \tag{30}$$

Note that $\bar{d} \leq d_{\max} \leq d$ and we have $\bar{d} = \mathbb{E}[\overline{D}]$, that is, the expected average degree of the network (on side 1). We need the following mild regularity assumption:

(A2) Consider the SBM model given by (4) and (5) with $n_1 \leq n_2$. Assume further that for some $\beta \geq 1$, and for all $t \in [k_1]$ and $s \in [k_2]$,

$$n_{1t} \geq \frac{n_1}{\beta k_1}, \quad n_{2s} \geq \frac{n_2}{\beta k_2}, \quad \text{and} \quad \left(\frac{n_2}{n_1}\right)^2 \max\{8\beta k_1, 8\beta k_2, 90\} \leq \bar{d} \leq \frac{1}{2}n_1. \tag{31}$$

Algorithm 1 Data-driven adjacency regularization

Input: Adjacency matrix $A \in \mathbb{R}^{n_1 \times n_2}$ and regularization parameter τ. (Default: $\tau = 3$)

Output: Regularized adjacency matrix $A_{\mathrm{re}} \in \mathbb{R}^{n_1 \times n_2}$.

1: Form the degree sequence $D_i = \sum_{j=1}^{n_2} A_{ij}$ for $i = 1, \ldots, n$ and the corresponding order statistics: $D_{(1)} \geq D_{(2)} \geq \cdots \geq D_{(n_1)}$.

2: Let $\overline{D} = \frac{1}{n_1} \sum_{i=1}^{n_1} D_i$ and $\alpha = \lfloor n_1/\overline{D} \rfloor$.

3: Set $\widehat{d}_1 = \tau D_{(\alpha)}$ and $\widehat{\mathcal{I}}_1 = \{i : D_i \geq \widehat{d}_1\}$.

4: Repeat Steps 1–3 on A^T to get $\widehat{\mathcal{I}}_2$ and $\widehat{d}_2$, and let D'_j be the corresponding degrees (i.e., column degrees in the original matrix A).

5: Set $[A_{\mathrm{re}}]_{ij} = A_{ij}$ for all $(i,j) \in \widehat{\mathcal{I}}_1^c \times \widehat{\mathcal{I}}_2^c$, and set $[A_{\mathrm{re}}]_{ij}$ arbitrarily when $i \in \widehat{\mathcal{I}}_1$ or $j \in \widehat{\mathcal{I}}_2$, but subject to: $\|[A_{\mathrm{re}}]_{i*}\|_1 \leq \widehat{d}_1$ for all $i \in \widehat{\mathcal{I}}_1$ and $\|[A_{\mathrm{re}}]_{*j}\|_1 \leq \widehat{d}_2$ for all $j \in \widehat{\mathcal{I}}_2$.

6: For example, $[A_{\mathrm{re}}]_{ij} = A_{ij} w_i w'_j$ for all $(i,j) \in [n_1] \times [n_2]$, where $w_i = \frac{\widehat{d}_1}{D_i} \wedge 1$ and $w'_j = \frac{\widehat{d}_2}{D'_j} \wedge 1$ satisfies the conditions in Step 5.

The next key lemma shows that $D_{(\alpha)}$ is a good proxy for $d_{\max}$ and d under a SBM:

Lemma 4 *Let $\alpha = \lfloor n_1/\overline{D} \rfloor$ where $\overline{D}$ is the average degree on side 1 as defined earlier. Under assumption (A2), the following hold, with probability at least $1 - 3e^{-n_1/80} - e^{-n_1/(\beta k_2)}$,*

(a) $d_{\max}/2 \leq D_{(\alpha)} \leq 3d_{\max}/2$.

(b) $\left| i : D_i > 3D_{(\alpha)} \right| \leq 10n_1/d$.

Conditions in (31) are quite mild: The first and second require that the clusters associated with the maximum degree are not too small. In the balanced case, where all clusters are of equal size, we have $n_{rt} = n_r/k_r$ (for all $t \in [k_r]$ and $r = 1, 2$) hence the condition is satisfied with $\beta = 1$. In general, β measures deviations from balancedness. Note that (A1) and the first two conditions in (31) do not imply each other. The last condition in (31) is satisfied if $\overline{d} \to \infty$ slower than n_1 and faster than $\max\{k_1, k_2\}$ (assuming $\beta = O(1)$ and $n_1 \asymp n_2$).

A result similar to Lemma 4 holds for the column degrees with appropriate modifications: Let $D'_j = \sum_i A_{ij}$ be the column degrees with expectation $d'_j = \sum_i p_{ij}$, and define the corresponding average $\overline{d}' = \frac{1}{n_2} \sum_j d'_j$ and $d'_{\max} = \max_j d'_j$. We, however, keep d as defined in (30).

Lemma 5 *Let $\alpha' = \lfloor n_2/\overline{D}' \rfloor$ where $\overline{D}'$ is the average degree on side 2 as defined earlier. Under assumption (A2), the following hold, with probability at least $1 - 3e^{-n_2/80} - e^{-n_1/(\beta k_1)}$,*

(a) $d'_{\max}/2 \leq D'_{(\alpha')} \leq 3d'_{\max}/2$.

(b) $\left| j : D'_j > 3D'_{(\alpha')} \right| \leq 10n_2/d$.

Equipped with Lemmas 4 and 5, we have the following for Algorithm 1:

Theorem 3 *Assume that $A \in \mathbb{R}^{n_1 \times n_2}$ is generated from the SBM model satisfying (A2). Then, for any $r \geq 1$, with probability at least $1 - n_2^{-r} - 6e^{-n_1/80} - 2e^{-n_1/\beta \max\{k_1,k_2\}}$, the regularized output A_{re} of Algorithm 1 (with $\tau = 3$) satisfies*

$$\|A_{re} - \mathbb{E}[A]\|_{op} \leq C_3 r^{3/2}\sqrt{d}, \tag{32}$$

where d is as in (30) and $C_3 > 0$ is a universal constant.

Proof [Proof of Theorem 3] Recall the definitions of $\widehat{\mathcal{I}}_r$ and $\widehat{d}_r$ ($r = 1, 2$) in Algorithm 1. On the union of the events described in Lemmas 4 and 5, both $|\widehat{\mathcal{I}}_1|$ and $|\widehat{\mathcal{I}}_2|$ are bounded by $10n_2/d$ (since $n_1 \leq n_2$), $\widehat{d}_1 = 3D_{(\alpha)} \leq 9d_{\max}/2$ and $\widehat{d}_2 = 3D'_{(\alpha')} \leq 9d'_{\max}/2$. Since $d_{\max} \leq n_2\|P\|_\infty = d$ and $d'_{\max} \leq n_1\|P\|_\infty \leq d$, we can apply Theorem 2 with $d' = 9d/2$ completing the proof. ∎

Theorem 3 thus provides the same concentration guarantee as in Theorem 2 without the knowledge of d. See Appendix A.2 for the proof of the two lemmas of this section.

Remark 9 (Comparison with existing work) *Results of the form (8) hold for A itself without any regularization if one further assumes that $d \gtrsim \log n_2$. These result are often derived for the symmetric case; see for example Tomozei and Massoulié (2010), Lei et al. (2015) and Chen and Xu (2016); or Bandeira et al. (2016) for the more general result with $d = \max_i \sum_j p_{ij}$. In order to break the $\log n_2$ barrier on the degrees, one has to resort to some form of regularization. When d is given, the general regularization for the adjacency matrix is to either to remove the high degree nodes as in Chin et al. (2015) or reduce their effect as in Le et al. (2017) and Theorem 2 above. In contrast, there is little work on data-driven regularization for the adjacency matrix. One such algorithm was investigated in Gao et al. (2017) for the special case of the SPBB model—discussed in Example 1 (Section 5)—under the assumption $a = O(b)$. The algorithm truncates the degrees to a multiple of the average degree, i.e., $\widehat{d}_1 = C\bar{D}$ for some large $C > 0$. A possible choice of $C = 5$ is given in Yun and Proutiere (2014a) assuming that the expected degrees of the nodes are similar.*

In contrast to existing results, our data-driven regularization holds for a general SBM and only requires the mild assumptions in (A2) while preserving the same upper bound (32) that holds with the knowledge of d. In particular, we do not require $\min_{k\ell} \Psi_{k\ell} \asymp \max_{k\ell} \Psi_{k\ell}$ (which is what $a = O(b)$ means in the SPBB model). Algorithm 1 enables effective and provable regularization without knowing any parameters.

5. Consistency results

We now state our various consistency results. The proofs are deferred to Appendix A.4. We start with a refinement of Theorem 1 for the specific algorithm SC-1 given in Algorithm 2.

Theorem 4 *Consider the spectral algorithm SC-1 given in Algorithm 2. Assume $k_1 = k =: \min\{k_1, k_2\}$, and for a sufficiently small $C > 0$,*

$$kd\,\sigma_k^{-2} \leq C(1 + \kappa)^{-2}.$$

Then, under the SBM model satisfying (A2), w.h.p.,

$$\overline{\mathrm{Mis}}(\mathcal{P}_\kappa(\hat{Z}_1), \bar{Z}_1) \lesssim (1 + \kappa)^2 \beta_1\left(\frac{d}{\sigma_k^2}\right).$$

Algorithm 2 SC-1

1: Apply regularization Algorithm 1 to A to obtain A_{re}.
2: Obtain the k-truncated SVD of A_{re} as $A_{\mathrm{re}}^{(k)} = \hat{Z}_1 \hat{\Sigma} \hat{Z}_2^T$. See (14).
3: Output an element of $\mathcal{P}_\kappa(\hat{Z}_1)$, i.e., a κ-approximate k-means solution for input $\hat{Z}_1$.

Algorithm 3 SC-RR

1: Apply regularization Algorithm 1 to A to obtain A_{re}.
2: Obtain the best rank k approximation of A_{re}, that is, $A_{\mathrm{re}}^{(k)} = \mathcal{T}_k(A_{\mathrm{re}})$. See (14).
3: Output an element of $\mathcal{P}_\kappa(A_{\mathrm{re}}^{(k)})$, i.e., a κ-approximate k-means solution for input $A_{\mathrm{re}}^{(k)}$.

---

where β_1 is given in (A1) *and d is defined in* (8).

One can take κ to be a fixed small constant say 1.5, since there are κ-approximate k-means algorithms for any $\kappa > 1$. In that case, $(1+\kappa)^2$ can be absorbed into other constants, and the bound in Theorem 4 is qualitatively similar to Theorem 1.

5.1. Reduced-rank SC

Theorem 4 implicitly assumes $\sigma_k > 0$, otherwise the bound is vacuous. This assumption is clearly violated if B is rank deficient (or equivalently, P has rank less than k). A variant of SC suggested in Yun and Proutiere (2014a) and Gao et al. (2018) can resolve this issue. The idea is to use the entire rank k approximation of A_{re}, and not just the singular vector matrix $\hat{Z}_1$, as the input to the k-means step. This approach, which we call reduced-rank SC, or SC-RR, is detailed in Algorithm 3. Recall the SV truncation operator $\mathcal{T}_k$ given in (14). It is well-known that $\mathcal{T}_k$ maps every matrix to its best rank-k approximation in Frobenius norm, i.e.,

$$\mathcal{T}_k(A_{\mathrm{re}}) = \min \left\{ \|R - A_{\mathrm{re}}\|_F : \ \mathrm{rank}(R) \leq k \right\}$$

with the approximation error satisfying

$$\|\mathcal{T}_k(A_{\mathrm{re}}) - A_{\mathrm{re}}\|_{\mathrm{op}} = \sigma_{k+1}(A_{\mathrm{re}}). \tag{33}$$

SC-RR uses this best rank-k approximation as a denoised version of A_{re} and runs a k-means algorithm on its rows. To analyze SC-RR, we need to replace bound (9) in Step 2 with an appropriate modification. The following lemma replaces Lemma 3 and provides the necessary bound in this case.

Lemma 6 *Let $A_{re}^{(k)} = \mathcal{T}_k(A_{re})$ be the k-truncated SVD of A_{re} and assume that the concentration bound* (8) *holds. Then,*

$$\|A_{re}^{(k)} - P\|_F \leq C\sqrt{8\,kd}. \tag{34}$$

Comparing with (9), we observe that (34) provides an improvement by removing the dependence on the singular value gap σ_k. However, we note that in terms of the relative

error, i.e., $\|A_{\text{re}}^{(k)} - P\|_F/\|P\|_F$ this may or may not be an improvement. There are cases where $\|P\|_F \approx \sqrt{k}\sigma_k$, in which case the relative error predicted by (34) is $O(\sqrt{d}/\sigma_k)$, similar to the relative error bound based on (9) (since $\|\bar{Z}_1 U_\psi Q\|_F = \sqrt{k}$); see Example 1 below.

Following through the three-step analysis of Section 3.1, with (9) replaced with (34), we obtain a qualitatively different bound on the misclassification error of Algorithm 3. The key is that center separation of P treated as a k-means matrix is different from that of $\bar{Z}_1 U_\psi Q$. Note that P is indeed a valid k-means matrix according to Definition (15); in fact, $P \in \mathbb{M}_{n_1,n_2}^{k_1}$. Similarly, $P^T \in \mathbb{M}_{n_2,n_1}^{k_2}$. Let us define

$$\Psi_{1,\wedge}^2 := \min_{(s,t):\, s \neq t} \sum_{\ell=1}^{k_2} \pi_{2\ell}(\Psi_{s\ell} - \Psi_{t\ell})^2, \tag{35}$$

$$\widetilde{\Psi}_{1,\wedge}^2 := \min_{(s,t):\, s \neq t} \left[\pi_{1t} \sum_{\ell=1}^{k_2} \pi_{2\ell}(\Psi_{s\ell} - \Psi_{t\ell})^2 \right]. \tag{36}$$

Theorem 5 *Consider the spectral algorithm* SC-RR *given in Algorithm 3. Assume that for a sufficiently small $C_1 > 0$,*

$$kd\,\widetilde{\Psi}_{1,\wedge}^{-2} \leq C_1(1+\kappa)^{-2}. \tag{37}$$

Then, under the SBM model satisfying (A2), with d as defined in (8), w.h.p.,

$$\overline{\text{Mis}}(\mathcal{P}_\kappa(A_{re}^{(k)}), P) \;\leq\; C_1^{-1}\,(1+\kappa)^2 \Big(\frac{kd}{\Psi_{1,\wedge}^2} \Big).$$

As is clear from the proof, one can take $C_1 = 1/(32C^2)$ where C is the constant in concentration bound (8). Condition (37) can be replaced with the stronger assumption

$$kd\,(\pi_{1,\wedge}\Psi_{1,\wedge}^2)^{-1} \leq C_1(1+\kappa)^{-2} \tag{38}$$

a where $\pi_{1,\wedge} := \min_{t \in [k_1]} \pi_{1t}$, since $\widetilde{\Psi}_{1,\wedge}^2 \geq \pi_{1,\wedge}\Psi_{1,\wedge}^2$.

Although the bounds of Theorems 4 and 5 are different, surprisingly, in the case of the planted partition model, they give the same result as the next example shows.

Example 1 (Planted partition model, symmetric case) *Let us consider the simplest symmetric SBM, the symmetric balanced planted partition (SBPP) model, and consider the consequences of Theorems 4 and 5 in this case. Recall that in the symmetric case we drop index r from k_r, n_r, n_{rj}, $\bar{N}_r$, β_r, $\Psi_{r,\wedge}$ and so on. SBPP is characterized by the following assumptions:*

$$\Psi = bE_k + (a-b)I_k, \;\; a \geq b, \quad \pi_j = n_j/n = \frac{1}{k}, \; \forall j \in [k].$$

Here, $E_k \in \mathbb{R}^{k \times k}$ is the all ones matrix and balanced *refers to all the communities being of equal size, leading to cluster proportions $\pi_j = 1/k$. In particular, $\beta = 1$, as defined in (A1). We have $\bar{B} = \bar{N}^{1/2}\Psi\bar{N}^{1/2} = \Psi/k$, recalling $\bar{N} = \text{diag}(\pi_j)$. Hence, the smallest singular value of $\bar{B}$ is $\sigma_k = (a-b)/k$. Theorem 4 gives the following result:*
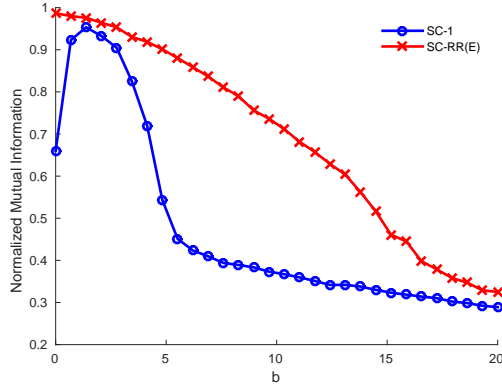
Figure 1: An example of the performance boost of SC-RR (or SC-RRE) relative to SC-1. The data is generated from the bipartite version of Example 2 with $n_2 = 2n_1 = 1000$, $k_1 = k_2 = 4$, $\pi_{r\ell} = n_r/k_r$ for all $\ell \in [k_r]$, $r = 1, 2$, and $\Psi = 2bE_4 + \text{diag}(16, 16, 16, 2)$ similar to (39). The key is the significant difference in the two smallest diagonal elements of Ψ. The plot shows the normalized mutual information (a measure of cluster quality) between the output of the two spectral clustering algorithms and the true clusters, as b varies. Only row clusters are considered. The plot shows a significant improvement for SC-RR(E) relative to SC-1 over a range of b. As b increases, the relative difference between Ψ_{33} and Ψ_{44} reduces and the model approaches that of Example 1, leading to similar performances for both algorithms as expected. It is interesting to note that the monotone nature of the performance of SC-RR(E) as a function of b and the non-monotone nature of that of SC-1 is reflected in the upper bounds (40) and (41).

Corollary 2 *Under the SBPP model, as long as $k^3 a/(a-b)^2$ is sufficiently small, SC-1 has average misclassification error of $O(k^2 a/(a-b)^2)$ with high probability.*

Now consider SC-RR. *Using definitions (35), we have $k\widetilde{\Psi}_\wedge^2 = \Psi_\wedge^2 = 2(a-b)^2/k$. Then, Theorem 5 gives the exact same result for* SC-RR:

Corollary 3 *Corollary 2 holds with* SC-1 *replaced with* SC-RR.

Results of Corollary 2 and 3 are consistency results as long as $k^2 a/(a-b)^2 = o(1)$. A typical example is when $k = O(1)$, $a = a_0 f_n$, $b = b_0 f_n$, $a_0 \asymp 1$ and $b_0 \asymp 1$ for some $f_n \to \infty$ as $n \to \infty$. Then, SC-1 *and* SC-RR *are both consistent at a rate $O(f_n^{-1})$.*

Let us now give an example where SC-1 and SC-RR behave differently.

Example 2 *Consider the symmetric balanced SBM, with*

$$\Psi = bE_k + \text{diag}(\alpha_1, \ldots, \alpha_k), \quad \pi_j = n_j/n = \frac{1}{k}, \ \forall j \in [k]. \tag{39}$$

As in Example 1, we have dropped the index r determining the side of network in the bipartite case. Let us assume that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k \geq 0$. We have

$$k\widetilde{\Psi}_\wedge^2 = \Psi_\wedge^2 = k^{-1} \min_{s \neq t} \sum_\ell (\Psi_{s\ell} - \Psi_{t\ell})^2 = k^{-1}(\alpha_{k-1}^2 + \alpha_k^2).$$

Thus, Theorem 5 gives the following: With ρ defined as follows:

$$\rho := k^2 \frac{\alpha_1 + b}{\alpha_{k-1}^2 + \alpha_k^2}, \tag{40}$$

as long as $k\rho$ is sufficiently small, SC-RR has average misclassification error $O(\rho)$ with high probability.

To determine the performance of SC-1, we need to estimate σ_k, the smallest singular value of $\bar{B} = \bar{N}^{1/2} \Psi \bar{N}^{1/2} = \Psi/k$. Since Ψ is obtained by a rank-one perturbation of a diagonal matrix, it is well-known that when $\{\alpha_t\}$ are distinct, the eigenvalues of Ψ are obtained by solving $\sum_{t=1}^{k} 1/(\alpha_t - \lambda) = -1/b$; the case where some of the $\{\alpha_t\}$ are repeated can be reasoned by the taking the limit of the general case. By plotting $\lambda \mapsto \sum_{t=1}^{k} 1/(\alpha_t - \lambda)$ and looking at the intersection with $\lambda \mapsto -1/b$, one can see that the smallest eigenvalue of Ψ, equivalently its smallest singular value, is in $[\alpha_k, \alpha_{k-1}]$, and can be made arbitrarily close to α_k by letting $b \to 0$. Letting $\alpha_k + \varepsilon_k(\alpha; b)$ denote this smallest singular value, we have $0 \le \varepsilon_k(\alpha; b) \to 0$ as $b \to 0$.

It follows that $\sigma_k = \sigma_k(\bar{B}) = k^{-1}(\alpha_k + \varepsilon_k(\alpha; b))$. Theorem 4 gives the following: With ρ defined as

$$\rho := k^2 \frac{\alpha_1 + b}{(\alpha_k + \varepsilon_k(\alpha; b))^2}, \tag{41}$$

as long as $k\rho$ is sufficiently small, SC-1 has average misclassification error $O(\rho)$ with high probability.

Comparing (41) with (40), the ratio of the two bounds is $(\alpha_{k-1}^2 + \alpha_k^2)/(\alpha_k + \varepsilon_k(\alpha; b))^2 \to 1 + (\alpha_{k-1}/\alpha_k)^2$ as $b \to 0$. This ratio could be arbitrarily large depending on the relative sizes of α_k and α_{k-1}. Thus, when the bounds give an accurate estimate of the misclassification rates of SC-1 and SC-RR, we observe that SC-RR has a clear advantage. This is empirically verified in Figure 1, for moderately dense cases. (In the very sparse case, the difference is not very much empirically.) In general, we expect SC-RR to perform better when there is a large gap between σ_k and σ_{k-1}, the two smallest nonzero singular values of $\bar{B}$.

Example 3 (Rank-deficient connectivity) *Consider an extreme case where Ψ is rank one: $\Psi = uv^T$ for some $u, v \in \mathbb{R}_+^k$, where again for simplicity we have assumed $k_1 = k_2 = k > 1$. Also assume $n_1 \asymp n_2 = n$ and $\pi_{rj} = 1/k$ for $j \in [k]$ and $r = 1, 2$, i.e., the clusters are balanced. In this case, $\bar{B} = \Psi/k$ and $\sigma_k = 0$, hence Theorem 4 does not provide any guarantees for SC-1. However, Theorem 3 is still valid. We have $k\widetilde{\Psi}_{1,\wedge}^2 = \Psi_{1,\wedge}^2 = k^{-1}\|v\|_2^2 \min_{s \ne t}(u_s - u_t)^2$ and $d \lesssim \|\Psi\|_\infty \le \|u\|_\infty \|v\|_\infty$. It follows from Theorem 3 that SC-RR has average misclassification rate bounded as*

$$O\left(\frac{k^2 \|u\|_\infty \|v\|_\infty}{\|v\|_2^2 \min_{s \ne t}(u_s - u_t)^2}\right)$$

whenever k times the above is sufficiently small. This is a consistency result assuming that the coordinates of u are different, all the elements of u and v are growing at the same rate and $k = O(1)$.

Algorithm 4 SC-RRE

1: Apply degree regularization Algorithm 1 to A to obtain A_{re}.
2: Obtain $A_{\mathrm{re}}^{(k)} = \hat{Z}_1 \hat{\Sigma} \hat{Z}_2^T$, the k-truncated SVD of A_{re}.
3: Output $\mathscr{K}(\hat{Z}_1 \hat{\Sigma})$ where $\mathscr{K}$ is an isometry-invariant κ-approximate k-means algorithm.

5.2. Efficient reduced-rank SC

The SC-RR algorithm discussed above has the disadvantage of running a k-means algorithm on vectors in $\mathbb{R}^n$ (the rows of $A_{\mathrm{re}}^{(k)}$, or in the ideal case the rows of P). We now introduce a variant of this algorithm that has the same performance as SC-RR in terms of misclassification rate, while computationally is as efficient as SC-1. This approach which we call efficient reduced-rank spectral clustering, SC-RRE, is detailed in Algorithm 4. The efficiency comes from running the k-means step on vectors in $\mathbb{R}^k$ which is usually a much smaller space than $\mathbb{R}^n$ ($k \ll n$ in applications).

For the k-means step in SC-RRE, we need a k-means (type) algorithm $\mathscr{K}$ that only uses the pairwise distances between the data points. We call such k-means algorithms *isometry-invariant*:

Definition 3 *A k-means (type) algorithm $\mathscr{K}$ is isometry-invariant if for any two matrices $X^{(r)} \in \mathbb{R}^{n \times d_r}, r = 1, 2$, with the same pairwise distances among points—i.e., $d(x_i^{(1)}, x_j^{(1)}) = d(x_i^{(2)}, x_j^{(2)})$ for all distinct $i, j \in [n]$, where $(x_i^{(r)})^T$ is the ith row of $X^{(r)}$—one has*

$$\overline{\mathrm{Mis}}(\mathscr{K}(X^{(1)}), \mathscr{K}(X^{(2)})) = 0.$$

Although the rows of $\mathscr{K}(X^{(1)})$ and $\mathscr{K}(X^{(2)})$ lie in spaces of possibly different dimensions, it still makes sense to talk about their relative misclassification rate, since this quantity only depends on the membership information of the k-means matrices and not their center information. We have implicitly assumed that $d(\cdot, \cdot)$ defines a family of distances over all Euclidean spaces $\mathbb{R}^d, d = 1, 2, \ldots$. This is obviously true for the common choice $d(x, y) = \|x - y\|_2$. If algorithm $\mathscr{K}$ is randomized, we assume that the same source of randomness is used (e.g., the same random initialization) when applying to either of the two cases $X^{(1)}$ and $X^{(2)}$.

The following result guarantees that SC-RRE behaves the same as SC-RR when one uses an isometry-invariant approximate k-means algorithm in the final step.

Theorem 6 *Consider the spectral algorithm SC-RRE given in Algorithm 4. Assume that for a sufficiently small $C_1 > 0$, (37) holds. Then, under the SBM model of Section 2, w.h.p.,*

$$\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1 \hat{\Sigma}), P) \;\leq\; C_1^{-1} (1 + \kappa)^2 \Big(\frac{kd}{\Psi_{1,\wedge}^2} \Big).$$

Remark 10 (Comparison with existing results) *The existing results hold under different assumptions. Table 1 provides a summary of some the recent results. The "p vs. q" denotes the assortative case where the diagonal entries of B are above p and off-diagonal entries are below q. The "e.v. dep" denotes whether the consistency result depends on the*

Table 1: Comparison of consistency results (cf. Remark 10).

	min. degree	p vs. q	$p = O(q)$	e.v. dep	given Ψ	k-means
Rohe et al. (2011)	$\Omega(n/\sqrt{\log n})$.	No.	Not needed.	Yes.	No.	$n \times k$.
Yun et al. (2014a)	Not needed.	Yes	Not needed.	No.	No.	$n \times n$.
Lei et al. (2015)	$\Omega(\log n)$.	No.	Not needed.	Yes.	No.	$n \times k$.
Chin et al. (2015)	Not needed.	Yes.	Required.	Yes.	Yes.	$n \times k$.
Gao et al. (2017)	Not needed.	Yes.	Required.	Yes.	No.	$n \times k$.
Gao et al. (2018)	Not needed.	Yes.	Required.	No.	No.	$n \times n$.
SC-1	Not needed.	No.	Not needed.	Yes.	No.	$n \times k$.
SC-RRE	Not needed.	No.	Not needed.	No.	No.	$n \times k$.

kth eigenvalue or singular value of B (or Ψ). The "k-means" column records the dimension of the matrix on which a k-means algorithm is applied. To allow for better comparison, let us consider a typical (special) case of the setting in this paper, where $n_1 = n_2$, $k_1 = k_2$, Ψ is symmetric and $d \asymp \Psi_{1,\wedge}$. Then all the spectral methods in Table 1 have misclassification rate guarantees that are polynomial in d^{-1}. General SBMs without assortative assumption (e.g., "p vs. q") were considered in earlier literature (Rohe et al., 2011; Lei et al., 2015). However, the theoretical guarantees were provided for sufficiently dense networks. The generalization to sparse networks is considered in Yun and Proutiere (2014a), Chin et al. (2015), Gao et al. (2017) and Gao et al. (2018) using some regularization on the adjacency matrix. However, these results only apply to assortative networks and with extra assumptions. Overall, the algorithms SC-1 and SC-RRE require less assumptions than any existing works. We also note that the guarantees of Theorems 5 and 6 in the context of a general SBM are new and have not appeared before (not even in the dense case).

5.3. Results in terms of mean parameters

One useful aspect of SC-RR(E) is that one can state its corresponding consistency result in terms of the *mean parameters* of the block model. Such results are useful when comparing to the optimal rates achievable in recovering the clusters. The row mean parameters of the SBM in Section 2 are defined as $\Lambda_{s\ell} := B_{s\ell}\, n_{2\ell}$ for $(s, \ell) \in [k_1] \times [k_2]$ which we collect in a matrix $\Lambda = (\Lambda_{s\ell}) \in \mathbb{R}^{k_1 \times k_2}$. To get an intuition for Λ note that

$$\mathbb{E}[AZ_2] = PZ_2 = Z_1 BN_2 = Z_1 \Lambda.$$

Each row of AZ_2 is obtained by summing the corresponding row of A over each of the column clusters to get a k_2 vector. In other words, the rows of AZ_2 are the sufficient statistics for estimating the row clusters, had we known the true column clusters. Note that $\mathbb{E}[(AZ_2)_{i*}] = z_{1i}^T \Lambda$, where the notation $(\cdot)_{i*}$ denotes the ith row of a matrix. In other words, we have $\mathbb{E}[(AZ_2)_{i*}] = \Lambda_{s*}$ if node i belongs to row cluster s. Let us define the minimum separation among these row mean parameters:

$$\Lambda_\wedge^2 := \min_{t \neq s} \|\Lambda_{s*} - \Lambda_{t*}\|^2. \tag{42}$$

We have the following corollary of Theorem 5 which is proved in Appendix A.4.

Corollary 4 *Assume that $\pi_{r,\wedge} := \min_t \pi_{rt} \geq (\beta_r k_r)^{-1}$ for $r = 1, 2$, and let $k = \min\{k_1, k_2\}$ and $\alpha = n_2/n_1$. Consider the spectral algorithm SC-RR given in Algorithm 3. Assume that for a sufficiently small $C_1 > 0$,*

$$\beta_1 \beta_2 \, k \, k_1 k_2 \, \alpha \frac{\|\Lambda\|_\infty}{\Lambda_\wedge^2} \leq C_1 (1 + \kappa)^{-2}. \tag{43}$$

Then, under the SBM model of Section 2, w.h.p.,

$$\overline{\mathrm{Mis}}(\mathcal{P}_\kappa(A_{re}^{(k)}), P) \;\leq\; C_1^{-1} \, (1 + \kappa)^2 \beta_2 \, k \, k_2 \, \alpha \frac{\|\Lambda\|_\infty}{\Lambda_\wedge^2}.$$

We note that the exact same result as Corollary 4 holds for SC-RRE assuming the k-means step uses an isometry invariant algorithm as discussed in Section 5. We refer to Zhou and Amini (2018) for an application of this result in constructing optimal clusterings in the bipartite setting.

6. Extensions

6.1. Clusters on one side only

The bipartite setting allows for the case where only one side has clusters. Assume that $A \sim \mathrm{Ber}(P)$ in the sense of (5) and, for example, only side 2 has k_2 clusters. Then we can model the problem as P having k_2 distinct columns. However, within columns we do not require any block constant structure, i.e., the k_2 distinct columns of P are general vectors in $[0, 1]^{n_1}$. This problem can be considered a special case of the SBM model discussed in Section 2 where $k_1 = n_1$: We recall that $P = \bar{Z}_1 \bar{B} \bar{Z}_2^T$ where $\bar{Z}_1$ is an orthogonal matrix of dimension $n_1 \times k_1$, hence $\bar{Z}_1 = I_{n_1}$. All the consistency results of the paper thus hold, where we set $k_1 = n_1$.

6.2. More clusters than rank

One of the unique features of the bipartite setting relative to the symmetric one is the possibility of having more clusters on one side of the network than the rank of the connectivity matrix. In the notation established so far, this is equivalent to $k_1 > k = \min\{k_1, k_2\}$. Let us first examine the performance of SC-1. Recall the SVD of $\bar{B} = U_\psi \Sigma V_\psi^T$, as given in Lemma 1. In contrast to the case $k_1 = k$, where the singular vector matrix U_ψ has no effect on the results (cf. Theorem 4), in the case $k_1 > k$, these singular vectors play a role. Recall that U_ψ is a $k_1 \times k$ orthogonal matrix, i.e., $U_\psi \in \mathbb{O}^{k_1 \times k}$. For a matrix $k_1 \times k_1$ matrix M, and index set $\mathcal{I} \subset [k_1]$, let $M_\mathcal{I}$ be the principal sub-matrix of M on indices $\mathcal{I} \times \mathcal{I}$. We assume the following *incoherence* condition:

$$\max_{\mathcal{I} \subset [k_1]: \, |\mathcal{I}|=2} \|(U_\psi U_\psi^T - I_{k_1})_\mathcal{I}\|_{\mathrm{op}} \;\leq\; 1 - \rho_1, \tag{44}$$

for some $\rho_1 \in (0, 1]$. Letting u_s^T be the sth row of U_ψ, for $s \in [k_1]$, we note that $(U_\psi U_\psi^T)_{st} = \langle u_s, u_t \rangle$, that is, $U_\psi U_\psi^T$ is the Gram matrix of the vectors u_s, $s \in [k_1]$. We have the following extension of Theorem 4:

Algorithm 5 SC-RRE for sub-Gaussian noise

1: Obtain $A^{(k)} = \hat{Z}_1 \hat{\Sigma} \hat{Z}_2^T$, the k-truncated SVD of A.
2: Output $\mathscr{K}(\hat{Z}_1 \hat{\Sigma})$ where $\mathscr{K}$ is an isometry-invariant κ-approximate k-means algorithm.

Theorem 7 *Consider the spectral algorithm* SC-1 *given in Algorithm 2. Assume that for a sufficiently small $C > 0$,*

$$kd\,\sigma_k^{-2} \leq C(1+\kappa)^{-2}\rho_1.$$

Then, under the SBM model of Section 2, w.h.p.,

$$\overline{\mathrm{Mis}}(\mathcal{P}_\kappa(\hat{Z}_1), \bar{Z}_1) \;\lesssim\; (1+\kappa)^2 \beta_1 \left(\frac{k}{\rho_1 k_1} \right) \left(\frac{d}{\sigma_k^2} \right). \tag{45}$$

The theorem is proven in Appendix A.5. The factor $k/(k_1\rho_1) = k_2/(k_1\rho_1)$ in (45) is the price one pays for the asymmetry of the number of communities, when applying SC-1. (Recall that $k := \min\{k_1, k_2\} = k_2$ by assumption.) Note that increasing k_1 decreases k/k_1, and at the same time, often increases ρ_1 since it is harder to have many nearly orthogonal unit vectors in low dimensions.

Remark 11 *It is interesting to note that in contrast to* SC-1*, the consistency results for* SC-RR(E) *do not need any modification for the case where the number of clusters is larger than the rank. In other words, the same Theorems 5 and 6 hold regardless of whether $k_1 = k$ or $k_1 > k$, though the difficulty of the latter case will be reflected implicitly via a reduction in $\Psi_{1,\wedge}^2$ and $\widetilde{\Psi}_{1,\wedge}^2$.*

6.3. General sub-Gaussian case

The analysis presented so far for network clustering problems can be extended to general sub-Gaussian similarity matrices. Consider a random matrix A with block constant mean

$$P := \mathbb{E}[A] = Z_1 B Z_2^T, \tag{46}$$

as defined in (4), and where $Z_r, r = 1, 2$ are again membership matrices. However, here A is not necessarily an adjacency matrix. We assume that A_{ij} are sub-Gaussian random variables independent across $(i,j) \in [n_1] \times [n_2]$ and let $\sigma := \max_{i,j} \|A_{ij} - \mathbb{E}A_{ij}\|_{\psi_2}$. We recall that a univariate random variable X is called sub-Gaussian if its sub-Gaussian norm is finite (Vershynin, 2018):

$$\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\} < \infty \tag{47}$$

Note that we do not assume A_{ij} and $A_{i'j'}$ to have the same distribution or the same sub-Gaussian norm even if $Z_{1i} = Z_{1i'}$ and $Z_{2j} = Z_{2j'}$.

Adapting Algorithm 4 to the general sub-Gaussian case, we have Algorithm 5 with the following performance guarantee:

Theorem 8 *Let $B_{1,\wedge}$ and $\widetilde{B}_{1,\wedge}$ be defined as in (35) with Ψ replaced with B and let*

$$\bar{n} := (n_1^{-1} + n_2^{-1})^{-1}.$$

Consider the spectral algorithm for sub-Gaussian noise given in Algorithm 5. Assume that for a sufficiently small $C_1 > 0$,

$$\frac{\sigma^2}{\bar{n}} k \, \widetilde{B}_{1,\wedge}^{-2} \leq C_1 (1 + \kappa)^{-2}. \tag{48}$$

Then, under the model defined in this section,

$$\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1\hat{\Sigma}), P) \ \leq \ C_1^{-1} (1 + \kappa)^2 \Big(\frac{\sigma^2}{\bar{n}} \frac{k}{B_{1,\wedge}^2} \Big).$$

with probability at least $1 - 2e^{-(n_1+n_2)}$.

See Appendix A.6 for the proof. Consider the typical case where the number of clusters k and the separation between them, $B_{1,\wedge}^2$, remains fixed as $\bar{n} \to \infty$. Then, $\sigma^2/\bar{n}$ plays the role of the signal-to-noise ratio (SNR) for the clustering problem and as long as $\sigma^2/\bar{n} = o(1)$, Algorithm 5 is consistent in recovering the clusters. Of course, there are other conditions under which we have consistency, e.g., cluster separation quantity $B_{1,\wedge}^2$ could go to zero, and k could grow as well (as $\bar{n} \to \infty$) and as long as $\sigma^2/\bar{n} = o(B_{1,\wedge}^2/k)$ the algorithm remains consistent.

6.4. Inhomogeneous random graphs

Up to now, we have stated consistency results for the SBM of Section 2. SBM is often criticized for having constant expected degree for nodes in the same community, in contrast to degree variation observed in real networks. (Although, SBM in the sparse regime $d = O(1)$ can still exhibit degree variation within a community, since the node degrees will be roughly $\mathrm{Poi}(d)$ distributed, showing little concentration around their expectations unless $d \to \infty$.) The popular remedy is to look at the degree-corrected block model (DC-SBM) (Karrer and Newman, 2011; Zhao et al., 2012; Gao et al., 2018). Instead, we consider the more general inhomogeneous random graph model (IRGM) (Söderberg, 2002; Bollobás et al., 2007) which might be more natural in practice, since it does not impose the somewhat parametric restrictions of DC-SBM on the mean matrix.

We argue that consistency results for the spectral clustering can be extended to a general inhomogeneous random graph (IRGM) model $A \sim \mathrm{Ber}(P)$, assuming that the mean matrix can be well-approximated by a block structure. Let us consider a scaling as before:

$$P := \frac{P^0}{\sqrt{n_1 n_2}}, \quad \text{and take} \quad d = \sqrt{\frac{n_2}{n_1}} \|P^0\|_\infty. \tag{49}$$

Note that $P \in [0,1]^{n_1 \times n_2}$ is a not assumed to have any block structure. However, we assume that there are membership matrices $Z_r \in \mathbb{H}_{n_r, k_r}$, $r = 1, 2$ such that P over the blocks defined by Z_1 and Z_2 is approximately constant. Let

$$\widetilde{B}_{st} := \frac{1}{n_{1s} n_{2t}} \sum_{ij} P_{ij} (Z_1)_{is} (Z_2)_{jt}, \quad s \in [k_1], \ t \in [k_2] \tag{50}$$

27

be the mean (or average) of P over these blocks, and let $\widetilde{B} = (\widetilde{B}_{st}) \in [0,1]^{k_1 \times k_2}$. Compactly, $\widetilde{B} = N_1^{-1} Z_1^T P Z_2 N_2^{-1}$, in the notation established in Section 2. We can define the SBM approximation of P as

$$\widetilde{P} := Z_1 \widetilde{B} Z_2^T = \bar{Z}_1 \bar{Z}_1^T P \bar{Z}_2 \bar{Z}_2^T = \Pi_1 P \Pi_2, \tag{51}$$

recalling $\bar{Z}_r = Z_r N_r^{-1/2}$ and introducing the notation $\Pi_r = \bar{Z}_r \bar{Z}_r^T$. Note that Π_r is a rank k_r projection matrix. According to (51), the map $P \mapsto \Pi_1 P \Pi_2$ takes a (mean) matrix to its SBM approximation relative to Z_1 and Z_2. We can thus define a similar approximation to P^0,

$$\widetilde{P}^0 := \Pi_1 P^0 \Pi_2, \quad \text{noting that} \quad \widetilde{P} = \frac{\widetilde{P}^0}{\sqrt{n_1 n_2}}.$$

In order to retain the qualitative nature of the consistency results, the deviation of each entry P_{ij}^0 from its block mean $(\widetilde{P}^0)_{ij}$ should not be much larger than $(\widetilde{P}^0)_{ij}^{1/2}$. In fact, we allow for a potentially larger deviation, assuming that:

$$\frac{1}{\sqrt{n_1 n_2}} \|P^0 - \widetilde{P}^0\|_F \le C_0 \sqrt{d}, \tag{52}$$

for some constant $C_0 > 0$ and a satisfying (49). The key is the following concentration result:

Proposition 2 *Assume that $A \sim \mathrm{Ber}(P)$ as in (5), with $n_1 \le n_2$, and let A_{re} be obtained from A by the regularization procedure in Theorem 2, with $d' = d$ as given in (49). Let $\widetilde{P}$ be as defined in (51), and assume that it satisfies (52). Then, with probability at least $1 - n_2^{-c}$,*

$$\|A_{re} - \widetilde{P}\|_{op} \le c_2 \sqrt{d}. \tag{53}$$

Proposition 2 provides the necessary concentration bound required in Step 1 of the analysis outlined in Section 3.1. In other words, bound (53) replaces (8) by guaranteeing a similar order of deviation for A_{re} around $\widetilde{P}$ instead of P. Thus, all the results of the paper follow under an IRGM with a mean matrix satisfying (52), with the modification that the connectivity matrix and all the related quantities (such as σ_k, $\Psi_{1,\wedge}^2$ and so on) are now based on $\widetilde{B}$, the connectivity matrix of the corresponding approximate SBM, as given in (50).

6.4.1. GRAPHON CLUSTERING

As a concrete example of the application of Proposition 2, let us consider a problem which we refer to as *graphon clustering*. For simplicity, consider the case $n_1 = n_2 = n$. Let $\rho_0 : [0,1]^2 \to \mathbb{R}_+$ be a bounded measurable function, and let $X_1, \ldots, X_n, Y_1, \ldots, Y_n \sim \mathrm{Unif}([0,1])$ be an i.i.d. sample. Assume that

$$A_{ij} \mid \{X_{i'}\}, \{Y_{j'}\} \ \sim \ \mathrm{Bern}\big(\rho_0(X_i, Y_j)/n\big), \tag{54}$$

independently over $i, j = 1, \ldots, n$. Consider two partitions of the unit interval: $\biguplus_{s=1}^{k_1} I_s = \biguplus_{t=1}^{k_2} J_t = [0,1]$ where I_s and J_t are (unknown) measurable sets. Let $\Psi \in [0,1]^{k_1 \times k_2}$ and

consider the block-constant function $\widetilde{\rho}_0$ on $[0,1]^2$ defined by

$$\widetilde{\rho}_0 := \sum_{s=1}^{k_1} \sum_{t=1}^{k_2} \Psi_{st} 1_{I_s \times J_t}, \tag{55}$$

where $1_{I_s \times J_t}(x,y) = 1_{I_s}(x) 1_{J_t}(y)$ is the indicator of the set $I_s \times J_t$. Whenever ρ_0 has a block constant approximation such as $\widetilde{\rho}_0$, we can consider I_s and J_t as defining implicit (true) clusters over nodes. More precisely, row node i belongs to community s if $1\{X_i \in I_s\}$ and similarly for the column labels. Note that none of $\{X_i\}$, $\{Y_j\}$, $\{I_s\}$ and $\{J_t\}$ are observed. One can ask whether we can still recover these implicit clusters given an instance of A.

Proposition 3 *Assume that there exists a function $\widetilde{\rho}_0$ of the form (55), such that*

$$\|\widetilde{\rho}_0 - \rho_0\|_{L^4} = o(\sqrt{d}) \tag{56}$$

Consider normalized mean matrices P^0 and $\widetilde{P}^0$ with entries, $P_{ij}^0 = \rho_0(X_i, Y_j)$ and $\widetilde{P}_{ij}^0 = \widetilde{\rho}_0(X_i, Y_j)$. Then, with high probability $\|P^0 - \widetilde{P}^0\|_F \leq n\sqrt{d}$.

Proposition 3 shows that a fourth moment bound of the form (56) is enough to guarantee (52) and as a consequence the result of Proposition 2. In order to apply the results of the paper, we require that (56) holds with $d = \|\rho_0\|_\infty$. Then, all the consistency results of the paper follow with Ψ replaced by that from (55), and regularization Algorithm 1 replaced with that of Theorem 2.

For example, with σ_k as defined in Section 2 (using Ψ from (55) in forming $\bar{B}$), Theorem 4 gives a misclassification rate at most $O(d/\sigma_k^2)$ for recovering the labels $1\{X_i \in I_s\}$ by the SC-1 algorithm. In typical cases, it is plausible to have $\sigma_k^2 \asymp d^2$ (cf. Remark 3), hence a misclassification rate of $O(1/d) = o(1)$ as $d \to \infty$. Proposition 3 is an illustrative example, and one can obtain other conditions by assuming more about the deviation $\widetilde{\rho}_0 - \rho_0$, such as boundedness.

Remark 12 *We note that graphon clustering problem considered above is different from what is typically called graphon estimation. In the latter problem, under a model of the form (54), one is interested in recovering the mean matrix $\mathbb{E}[A]$ or ρ_0 in the MSE sense. This problem has been studied extensively in recent years, often under the assumption of smoothness of ρ_0, for maximum likelihood SBM approximation (Airoldi et al., 2013; Olhede and Wolfe, 2014; Gao et al., 2015; Klopp et al., 2017) and spectral truncation (Xu, 2017). The graphon clustering problem, as far as we know, has not been considered before and is concerned with recovering the underlying clusters, assuming that such true clusters exist. We only need the existence of a block constant approximation $\widetilde{\rho}_0$, over the true clusters (of the form (55)) that satisfies $\|\widetilde{\rho}_0 - \rho_0\|_{L^4} = o(\sqrt{d})$ for $d = \|\widetilde{\rho}_0\|_\infty$. Then our results implicitly imply that the underlying clusters are identifiable and consistently recovered by spectral approaches. Note that we do not impose any explicit smoothness assumption on ρ and there is no lower bound requirement on d (such as $d = \Omega(\log n)$ in Xu (2017)).*
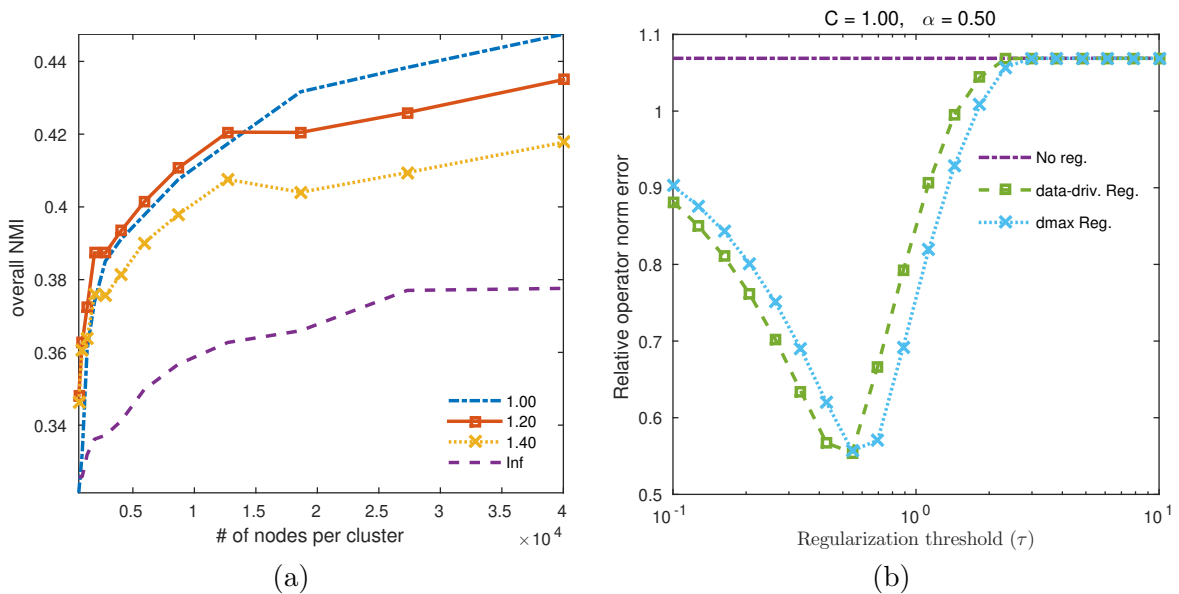
Figure 2: (a) NMI plots for SC-RRE algorithm with degree regularization Algorithm 1 for various values of $\tau = 1, 1.2, 1.4$ and ∞. (b) Relative operator norm error between the adjacency matrix and its expectation, with data-driven and oracle ($d_{\max}$) regularization as well as no regularization.

7. Simulations

We now present some simulation results showing the performance of the data-driven regularization of Section 4.1. We sample from the bipartite SBM model with connectivity matrix

$$
B = \sqrt{\frac{\log(n_1 n_2)}{n_1 n_2}} B_0, \quad \text{where} \quad B_0 = \frac{1}{2} \begin{bmatrix} 6 & 1 & 1 & 1 \\ 1 & 6 & 1 & 1 \\ 1 & 1 & 6 & 1 \end{bmatrix}, \tag{57}
$$

for which $\Psi = \sqrt{\log(n_1 n_2)} B_0$ as in (4). We let $n_1 = n_0 k_1$ and $n_2 = n_0 k_2$, and we vary n_0. Note that $k_1 = 3$ and $k_2 = 4$. We measure the performance using the normalized mutual information (NMI) between the true and estimated clusters. The NMI belongs to the interval $[0, 1]$ and is monotonically increasing with clustering accuracy. We consider Algorithm 1 with regularization parameter $\tau = 1, 1.2, 1.4$ and ∞, where $\tau = \infty$ corresponds to no regularization. Although we have established theoretical guarantees for $\tau = 3$, this constant is not optimal and any scalar $\asymp 1$ might perform well.

In this model, the key parameter $d = \sqrt{n_2/n_1} \|\Psi\|_\infty \sim \sqrt{\log n_0}$ as $n_0 \to \infty$. In other words, the maximum expected degree of the network scales as $\sqrt{\log n_0}$ which is enough for the consistency of spectral clustering; in fact, results in Section 5 predict a misclassification rate of $O(d^{-1}) = O\big((\log n_0)^{-1/2}\big)$ for various spectral algorithms discussed in this paper.

Figure 2(a) shows the NMI plots as a function of n_0 for the SC-RRE algorithm with the regularization scheme of Algorithm 1. In the k-means step, we have used `kmeans++` which as described in Remark 8 satisfies the approximation property of Section 3.3 with $\kappa = O(\log(k_1 \wedge k_2)) = O(1)$ in this case. The results are averaged over 15 replicates. The

plots clearly show that the regularization considerably boosts the performance of spectral clustering for model (57).

Figure 2(b) shows the relative operator norm error between the (regularized) adjacency matrix and its expectation, i.e., $\|A_{\mathrm{re}} - \mathbb{E}A\|_{\mathrm{op}}/\|\mathbb{E}A\|_{\mathrm{op}}$ with and without regularization ($A_{\mathrm{re}} = A$). The plots correspond to the same SBM model with $n_0 = 500$ (and the results are averaged over 3 replicates). For the regularization, we consider both the oracle where the degrees are truncated to $\tau d_{\max}$ and $\tau d'_{\max}$ for the rows and columns (see Section 4.1 for the definitions of these quantities) as well as the data-driven one that truncates as in Algorithm 1. The plots show the relative error as a function of τ and we see that the regularization clearly improves the concentration. We also note that the behavior of the data-driven truncation closely follows that of the oracle as predicted by Theorem 3.

The code is available at (Zhou and Amini, 2018).

Acknowledgement

Appendix A. Proofs

A.1. Proofs of Section 3.2

Proof [Proof of Lemma 3] Let $\bar{W}$ and $\widehat{W}$ be the W of Lemma 2(a) for $P^\dagger$ and $A_{\mathrm{re}}^\dagger$, respectively. Let us also write $\bar{W}_1$ and $\widehat{W}_1$ for the $(n_1 + n_2) \times k$ matrices obtained by taking the submatrices of $\bar{W}$ and $\widehat{W}$ on columns $1, \ldots, k$. We have

$$\bar{W}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} \bar{Z}_1 U_\psi \\ \bar{Z}_2 V_\psi \end{pmatrix}, \quad \widehat{W}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} \hat{Z}_1 \\ \hat{Z}_2 \end{pmatrix}.$$

Note that $\bar{W}_1, \widehat{W}_1 \in \mathbb{O}^{(n_1 + n_2) \times k}$. Let $\Pi_{\bar{W}_1}$ be the (orthogonal) projection operator, projecting onto $\mathrm{Im}(\bar{W}_1)$, i.e., the column span of $\bar{W}_1$, and similarly for $\Pi_{\widehat{W}_1}$. We have

$$\begin{aligned}
\|\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}\|_{\mathrm{op}} &\leq \frac{2}{2\sigma_k} \|A_{\mathrm{re}}^\dagger - P^\dagger\|_{\mathrm{op}} && \text{(Symmetric DK and Lemma 2(d))} \\
&= \frac{1}{\sigma_K} \|(A_{\mathrm{re}} - P)^\dagger\|_{\mathrm{op}} && \text{(Linearity of dilation)} \\
&= \frac{1}{\sigma_K} \|A_{\mathrm{re}} - P\|_{\mathrm{op}} && \text{(Lemma 2(b)).}
\end{aligned}$$

The next step is to translate the operator norm bound on spectral projections into a Frobenius bound. The key here is the bound on the rank of spectral deviations which leads to a $\sqrt{k}$ scaling as opposed to $\sqrt{n_1 + n_2}$, when translating from operator norm to Frobenius:

$$\begin{aligned}
\min_{Q \in \mathbb{O}^{k \times k}} \|\widehat{W}_1 - \bar{W}_1 Q\|_F &\leq \|\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}\|_F && \text{(By Lemma 8 in Appendix C)} \\
&\leq \sqrt{2k} \, \|\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}\|_{\mathrm{op}} && \left(\mathrm{rank}(\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}) \leq 2k \right) \\
&\leq \frac{\sqrt{2k}}{\sigma_k} \|A_{\mathrm{re}} - P\|_{\mathrm{op}}.
\end{aligned}$$

Since $2\|\widehat{W}_1 - \bar{W}_1 Q\|_F^2 = \|\hat{Z}_1 - \bar{Z}_1 U_\psi Q\|_F^2 + \|\hat{Z}_2 - \bar{Z}_2 V_\psi Q\|_F^2$, we obtain the desired result after combining with (8). ■

A.2. Proofs of Section 4

Let us start with a relatively well-known concentration inequality:

Proposition 4 (Prokhorov) *Let* $S = \sum_i X_i$ *for independent centered variables* $\{X_i\}$, *each bounded by* $c < \infty$ *in absolute value a.s. and suppose* $v \geq \sum_i \mathbb{E} X_i^2$, *then*

$$\mathbb{P}(S > vt) \leq \exp[-v h_c(t)], \quad t \geq 0, \quad where \ h_c(t) := \frac{3}{4c} t \log\left(1 + \frac{2c}{3}t\right). \tag{58}$$

Same bound holds for $\mathbb{P}(S < -vt)$.

We often apply this result with $c = 1$. We note that for any $u \geq \frac{2}{3}\alpha$ any $\alpha > 0$,

$$h_1(\alpha u) \geq \frac{3}{4}\alpha u \log\left(\frac{2}{3}\alpha u\right) \geq \frac{3}{4}\alpha u \log u^2 \geq \frac{3}{2}\alpha u \log u. \tag{59}$$

Proof [Proof of Lemma 4] We first note that

$$\mathbb{P}(D_i - d_i > vt) \leq \exp(-v h_1(t)), \quad \forall v \geq d_i, \ t \geq 0, \tag{60}$$

$$\mathbb{P}(D_i - d_i < -vt) \leq \exp(-v h_1(t)), \quad \forall v \geq d_i, \ t \geq 0. \tag{61}$$

Lower bound. Let $Y_i = 1\{D_i < d_{\max}/2\}$. Without loss of generality, assume that cluster 1 achieves the maximum in the definition of $d_{\max}$ (i.e., $t^* = 1$), and fix i such that $z_{1i} = 1$. We have $d_i = d_{\max}$, hence applying (61) with $v = d_i = d_{\max}$ and $t = 1/2$,

$$\mathbb{P}(D_i < d_{\max}/2) \leq \exp(-d_{\max}/10),$$

using $h_1(1/2) = 0.2157 > 1/10$. Let $q := \exp(-d_{\max}/10)$ so that $\mathbb{E} Y_i \leq q$. By assumption $d_{\max} \geq \bar{d}$ is sufficiently large that $q \leq 1/4$. We then have

$$
\begin{aligned}
\mathbb{P}(D_{(\lfloor n_1/\bar{D}\rfloor)} < d_{\max}/2) &\leq \mathbb{P}\left(\sum_{i=1}^{n_1} 1\{D_i \geq d_{\max}/2\} \leq n_1/\bar{D}\right) \leq \mathbb{P}\left(\sum_{i:z_{1i}=1} 1\{D_i \geq d_{\max}/2\} \leq n_1/\bar{D}\right) \\
&= \mathbb{P}\left(\sum_{i:z_{1i}=1} Y_i > n_{11} - n_1/\bar{D}\right) \\
&\leq \mathbb{P}\left(\sum_{i:z_{1i}=1} Y_i > n_{11} - 2n_1/\bar{d}\right) + \mathbb{P}\left(\bar{D} < \frac{\bar{d}}{2}\right) =: T_1 + T_2
\end{aligned}
$$

For the first term we have

$$T_1 \leq \mathbb{P}\left(\sum_{i:z_{1i}=1} (Y_i - \mathbb{E} Y_i) > n_{11}(1 - q) - 2n_1/\bar{d}\right).$$

By assumptions $n_{11} \geq n_1/(\beta k_1)$ and $\beta k_1/\bar{d} \leq 1/8$, we have $n_{11}(1-q) - 2n_1/\bar{d} \geq n_{11}/2$. Taking $c = 1$, $v = n_{11}q$, $t = 1/(2q)$ in Proposition 4, we have $vt = n_{11}/2$ and $t \geq 2$

$$T_1 \leq \mathbb{P}\Big(\sum_{i:\, z_{1i}=1} (Y_i - \mathbb{E}Y_i) > n_{11}/2 \Big) \leq \exp\Big(-n_{11}\, q\, h_1\Big(\frac{1}{2q}\Big)\Big)$$

$$\leq \exp\big(-c_1\, n_{11} d_{\max}\big) \leq e^{-3n_1/5}$$

where we have applied (59) with $\alpha = 1/2$ and $u = 1/q = e^{d_{\max}/10}$ to get

$$q\, h_1\Big(\frac{1}{2q}\Big) \geq q\frac{3}{2}\frac{1}{2q}\log(1/q) = \frac{3}{40}d_{\max}$$

and $n_{11}d_{\max} \geq n_1 d_{\max}/(\beta k_1) \geq 8n_1$ since $\beta k_1/d_{\max} \leq \beta k_1/\bar{d} \leq 1/8$ by assumption.

For the second term T_2, we note that $n_1(\bar{D} - \bar{d}) = \sum_{i,j}(A_{ij} - \mathbb{E}A_{ij})$. Applying Proposition 4 with $v = n_1\bar{d}$ and $t = 1/2$, we have

$$T_2 = \mathbb{P}\Big(n_1(\bar{D} - \bar{d}) < -n_1\bar{d}/2\Big) \leq \exp(-n_1\bar{d}/10) \leq e^{-n_1/5}. \tag{62}$$

using $h_1(1/2) \geq 1/10$ and $\bar{d} \geq 2$.

Upper bound. For any $\in [n_1]$, applying (60) with $v = d_{\max} \geq d_i$, and $t = 1/2$, we have

$$\mathbb{P}(D_i > 3d_{\max}/2) \leq \exp(-d_{\max}/10).$$

Let $Z_i = 1\{D_i > 3d_{\max}/2\}$ and note that $\mathbb{E}Z_i \leq q = \exp(-d_{\max}/10)$, as in the case of the lower bound. Then,

$$\mathbb{P}\Big(D_{(\lfloor n_1/\bar{D}\rfloor)} > \frac{3}{2}d_{\max}\Big) \leq \mathbb{P}\Big(\sum_{i=1}^{n_1} Z_i \geq \frac{n_1}{\bar{D}} - 1\Big)$$

$$\leq \mathbb{P}\Big(\sum_{i=1}^{n_1} Z_i \geq \frac{2n_1}{3\bar{d}} - 1\Big) + \mathbb{P}\Big(\bar{D} > \frac{3\bar{d}}{2}\Big) =: T_1' + T_2'.$$

Using assumption $n_1/\bar{d} \geq 2$, we have $2n_1/(3\bar{d}) - 1 \geq n_1/(6\bar{d}) \geq n_1/(6d_{\max})$, hence

$$T_1' \leq \mathbb{P}\Big(\sum_{i=1}^{n_1}(Z_i - \mathbb{E}Z_i) \geq \frac{n_1}{6d_{\max}} - n_1 q\Big) =: T_1'' \tag{63}$$

We claim that if $d_{\max} \geq 1$, then for any $\gamma \geq 1$,

$$f(\gamma) := \mathbb{P}\Big(\sum_{i=1}^{n_1}(Z_i - \mathbb{E}Z_i) \geq \frac{n_1}{\gamma d_{\max}}\Big) \leq e^{-3n_1/(20\gamma)}. \tag{64}$$

To see this, taking $c = 1$, $v = n_1 q$, $t = 1/(\gamma q d_{\max})$ in Proposition 4, we have

$$f(\gamma) \leq \exp\Big(-n_1 q\, h_1\Big(\frac{1}{\gamma q d_{\max}}\Big)\Big).$$

33

Applying (59) with $\alpha = 1/(\gamma d_{\max})$ and $u = 1/q = e^{d_{\max}/10}$, noting that $u \geq (2/3)\alpha$,

$$q\, h_1\Big(\frac{1}{\gamma q d_{\max}}\Big) \geq q\frac{3}{2}\frac{1}{\gamma q d_{\max}}\log(1/q) = \frac{3}{20\gamma}$$

showing (64).

Going back to bounding T_1'', for sufficiently large $d_{\max}$, we have $1/(6d_{\max}) - q \geq 1/(12d_{\max})$. Hence, we can apply (64) with $\gamma = 12$ to conclude, $T_1'' \leq e^{-n_1/80}$. For T_2', by the same argument as in (62), we have $T_2 \leq e^{-n_1/5}$.

Proof of part (b). From the lower bound, we have that $3D_{(\lfloor n_1/\bar{D} \rfloor)} > \frac{3}{2}d_{\max}$ on an event $\mathcal{A}$ with probability $\mathbb{P}(\mathcal{A}) \geq 1 - 2e^{-n_1/80}$. Then,

$$W_i := 1\big\{D_i > 3D_{(\lfloor n_1/\bar{D} \rfloor)}\big\} \leq 1\{D_i > 3d_{\max}/2\} = Z_i, \quad \text{on } \mathcal{A}.$$

Hence, $\big|i : D_i > 3D_{(\lfloor n_1/\bar{D} \rfloor)}\big| = \sum_i W_i$ and

$$\mathbb{P}\Big(\Big\{\sum_{i=1}^{n_1} W_i > \frac{10n_1}{d}\Big\} \cap \mathcal{A}\Big) \leq \mathbb{P}\Big(\sum_{i=1}^{n_1} Z_i > \frac{10n_1}{d}\Big) = \mathbb{P}\Big(\sum_{i=1}^{n_1}(Z_i - \mathbb{E}Z_i) > \frac{10n_1}{d} - n_1 q\Big).$$

Since $d_{\max} \geq n_2\|P\|_\infty/(\beta k_2) = d/(\beta k_2)$, and $\beta k_2 \leq \bar{d}/8 \leq d_{\max}/8$ we have $d_{\max}^2 \geq 8d$. Recalling that $q = e^{-d_{\max}/10}$

$$qd = (qd_{\max}^2)(d/d_{\max}^2) \leq 1/8. \tag{65}$$

for $d_{\max} \geq 90$. We have

$$\frac{10n_1}{d} - n_1 q \geq \frac{9n_1}{d} \geq \frac{9n_1}{d_{\max}}\frac{1}{\beta k_2}$$

hence we can apply (64) with $\gamma = \beta k_2/9$ to obtain

$$\mathbb{P}\Big(\sum_{i=1}^{n_1} W_i > \frac{10n_1}{d}\Big) \leq \exp\Big(-\frac{27n_1}{20\beta k_2}\Big).$$

The proof is complete. ∎

Proof [Proof of Lemma 5] The proof of part (a) follows exactly as in the case of row degrees establishing the result with probability at least $1 - 3e^{-n_2/80}$. For part (b), we have by the same argument as in the case of row degrees

$$\mathbb{P}\Big(\big|j : D_j' > 3D_{(\lfloor n_2/\bar{D}' \rfloor)}'\big| > \frac{10n_2}{d}\Big) \leq \mathbb{P}\Big(\sum_{j=1}^{n_2}(Z_j' - \mathbb{E}Z_j') > \frac{10n_2}{d} - n_2 q'\Big) + \mathbb{P}(\mathcal{A}'^c)$$

$$:= T_3 + \mathbb{P}(\mathcal{A}'^c)$$

where $\mathcal{A}'$ is defined similar to $\mathcal{A}$ for column variables and is controlled similarly. We also have $q' := \exp(-d_{\max}'/10)$.

To simplify notation, let $\alpha := n_2/n_1 \geq 1$. We have $d'_{\max} \geq n_1 \|P\|_\infty/(\beta k_1) = d/(\alpha\beta k_1)$, recalling $d = n_2 \|P\|_\infty$. By assumption, $\bar{d} \geq 8\beta k_1 \alpha^2$. Since $n_1 \bar{d} = \sum_{ij} p_{ij} = n_2 \bar{d}'$, that is, $\bar{d} = \alpha \bar{d}'$, we obtain $\beta k_1 \leq \bar{d}/(8\alpha^2) = \bar{d}'/(8\alpha) \leq d'_{\max}/(8\alpha)$. It follows that $(d'_{\max})^2 \geq 8d$ which is similar to what we had for the rows; hence, $q'd \leq 1/8$ as long as $d'_{\max} \geq 90$ (see (65)), which is true since $d'_{\max} \geq \bar{d}' = \bar{d}/\alpha \geq \bar{d}/\alpha^2 \geq 90$ by assumption. Then,

$$\frac{10n_2}{d} - n_2 q' \geq \frac{9n_2}{d} \geq \frac{9n_2}{d'_{\max}} \frac{1}{\alpha\beta k_1}.$$

Therefore, applying the column counterpart of (64) with $\gamma = \alpha\beta k_1/9$,

$$T_3 \leq \exp\left(-\frac{27n_2}{20\alpha\beta k_1}\right) = \exp\left(-\frac{27n_1}{20\beta k_1}\right).$$

The proof is complete. ∎

A.3. Proofs of Section 3.3

Proof [Proof of Corollary 1] Using (21), we have $d_F(X^*, \widetilde{X}) \leq (1+\kappa)\varepsilon$ for any $\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})$. We now apply Proposition 1 to X^* and $\widetilde{X}$, both k-means matrices, with $(1 + \kappa)\varepsilon$ in place of ε. ∎

Proof [Proof of Proposition 1] The proof follows the argument in (Lei et al., 2015, Lemma 5.3) which is further attributed to Jin (2015). Let $\mathcal{C}_r$ denote the rth cluster of X, having center $q_r = q_r(X)$. We have $|\mathcal{C}_r| = n_r$. Let x_i^T and $\widetilde{x}_i^T$ be the ith row of X and $\widetilde{X}$, respectively, and let

$$T_r := \{i \in \mathcal{C}_r : d(\widetilde{x}_i, q_r) < c_r \delta_r\} = \{i \in \mathcal{C}_r : d(\widetilde{x}_i, x_i) < c_r \delta_r\}$$

using $x_i = q_r$ for all $i \in \mathcal{C}_r$ which holds by definition. Let $S_r = \mathcal{C}_r \setminus T_r$. Then,

$$|S_r| c_r^2 \delta_r^2 \leq \sum_{i \in S_r} d(\widetilde{x}_i, x_i)^2 \leq \varepsilon^2 \implies \frac{|S_r|}{|\mathcal{C}_r|} \leq \frac{c_r^{-2}\varepsilon^2}{n_r \delta_r^2} < 1. \tag{66}$$

where we have used assumption (b). It follows that S_r is a proper subset of $\mathcal{C}_r$, that is, T_r is nonempty for all $r \in [k]$.

Next, we argue that if two elements belong to different $T_r, r \in [k]$, they have different labels according to $\widetilde{X}$. That is, $i \in T_r$, $j \in T_\ell$ for $r \neq \ell$ implies $\widetilde{x}_i \neq \widetilde{x}_j$. Assume otherwise, that is, $\widetilde{x}_i = \widetilde{x}_j$. Then, by triangle inequality and $c_r + c_\ell \leq 1$,

$$d(q_k, q_\ell) \leq d(q_k, \widetilde{x}_i) + d(q_\ell, \widetilde{x}_j) < c_r \delta_r + c_\ell \delta_\ell \leq \max\{\delta_r, \delta_\ell\}$$

contradicting (23). This shows that $\widetilde{X}$ has at least k labels, since all T_r are nonempty, hence exactly k labels, since $\widetilde{X} \in \mathbb{M}_{n,m}^k$ by assumption.

Finally, we argue that if two elements belong to the same T_r, they have the same label according to $\widetilde{X}$. This immediately follows from the previous step since otherwise there will

be at least $k+1$ labels. Thus, we have shown that, for all $r \in [k]$, the labels in each T_r are in the same cluster according to both X and $\widetilde{X}$, that is, they are correctly classified. The misclassification rate over cluster $\mathcal{C}_r$ is then $\leq |S_r|/|\mathcal{C}_r|$ which establishes the result in view of (66). ∎

A.4. Proofs of Section 5

Proof [Proof of Theorem 4] Going through the three-step plan of analysis in Section 3, we observe that (8) holds for A_{re} by Theorem 2, and (9) holds by Lemma 3. We only need to verify conditions of Corollary 1, so that κ-approximate k-means operator $\mathcal{P}_\kappa$ satisfies bound (10) of the k-means step. As in the proof of Theorem 1, $X^* = \bar{Z}_1 O \in \mathbb{R}^{n_1 \times k}$, where $\bar{Z}_1 \in \mathbb{O}^{n_1 \times k_1}$ and $O := U_\psi Q \in \mathbb{O}^{k_1 \times k}$. Clearly, X^* has exactly k distinct rows (recalling $k = k_1$). Furthermore, using the calculation in the proof of Theorem 1,

$$n_{1t}\,\delta_t^2 = n_{1t} \min_{s:\, s \neq t}\left(n_{1t}^{-1} + n_{1s}^{-1}\right) = \min_{s:\, s \neq t}\left(1 + \frac{n_{1t}}{n_{1s}}\right) \geq 1.$$

Recalling that $\varepsilon^2 = C_2^2 k d/\sigma_k^2$, as long as

$$4(1+\kappa)^2 \varepsilon^2 = 4C_2^2(1+\kappa)^2\, kd/\sigma_k^2 < 1 \leq n_{1t}\,\delta_t^2$$

condition (b) of Corollary 1 holds and $\mathcal{P}_\kappa$ satisfies (10) with $c = 4(1+\kappa)^2$ as in (26). The rest of the proof follows as in Theorem 1. ∎

Proof [Proof of Lemma 6] Throughout the proof, let $\|\cdot\| = \|\cdot\|_{\mathrm{op}}$ be the operator norm. Recall that $P = \mathbb{E}A$ is the mean matrix itself, and let $\Delta_{\mathrm{re}} := A_{\mathrm{re}} - P$. By Weyl's theorem on the perturbation of singular values, $|\sigma_i(A_{\mathrm{re}}) - \sigma_i(P)| \leq \|\Delta_{\mathrm{re}}\|$ for all i. Since $\sigma_{k+1}(P) = 0$ (see (7)), we have $\sigma_{k+1}(A_{\mathrm{re}}) \leq \|\Delta_{\mathrm{re}}\|$, hence

$$
\begin{aligned}
\|A_{\mathrm{re}}^{(k)} - P\| &\leq \|A_{\mathrm{re}}^{(k)} - A_{\mathrm{re}}\| + \|\Delta_{\mathrm{re}}\| && \text{(triangle inequality)} \\
&= \sigma_{k+1}(A_{\mathrm{re}}) + \|\Delta_{\mathrm{re}}\| && \text{(by (33))} \\
&\leq 2\|\Delta_{\mathrm{re}}\| && \text{(Weyl's theorem).}
\end{aligned}
$$

Thus, in terms of the operator norm, we lose at most a constant in going from A_{re} to $A_{\mathrm{re}}^{(k)}$. However, we gain a lot in Frobenious norm deviation. Since A_{re} is full-rank in general, the best bound on Δ_{re} based on its operator norm is $\|\Delta_{\mathrm{re}}\|_F \leq \sqrt{n_\wedge}\,\|\Delta_{\mathrm{re}}\|$ where $n_\wedge = \min\{n_1, n_2\}$. On the other hand, since $A_{\mathrm{re}}^{(k)} - P$ is of rank $\leq 2k$, we get

$$\|A_{\mathrm{re}}^{(k)} - P\|_F \leq \sqrt{2k}\|A_{\mathrm{re}}^{(k)} - P\| \leq 2\sqrt{2k}\|\Delta_{\mathrm{re}}\|.$$

Combining with (8), that is, $\|\Delta_{\mathrm{re}}\| \leq C\sqrt{d}$, we have the result. ∎

Proof [Proof of Theorem 5] We only need to calculate $\delta(P) = \delta_\wedge(P)$ the minimum center separation of P viewed as an element of $\mathbb{M}_{n_1, n_2}^{k_1}$. Recall that

$$P = \bar{Z}_1 \bar{B} \bar{Z}_2^T = Z_1 N_1^{-1/2} (\bar{N}_1^{1/2}\Psi \bar{N}_1^{1/2})\bar{Z}_2^T = n_1^{-1/2} Z_1 \Psi \bar{N}_1^{1/2}\bar{Z}_2^T.$$

Let e_s be the sth standard basis vector of $\mathbb{R}^{k_1}$. Unique rows of P are $q_s^T := n_1^{-1/2} e_s^T (\Psi \bar{N}_2^{1/2}) \bar{Z}_2^T$ for $s \in [k_1]$. We have

$$\|q_s - q_t\|_2^2 = n_1^{-1} \|\bar{Z}_2 \bar{N}_2^{1/2} \Psi^T (e_s - e_t)\|_2^2$$

$$= n_1^{-1} \|\bar{N}_2^{1/2} \Psi^T (e_s - e_t)\|_2^2 = n_1^{-1} \sum_{\ell=1}^{k_2} \pi_{2\ell} (\Psi_{s\ell} - \Psi_{t\ell})^2.$$

It follows that $\delta^2(P) = \min_{t \neq s} \|q_s - q_t\|_2^2 = n_1^{-1} \Psi_{1,\wedge}^2$. We apply Corollary 1, with $X^* = P$ and $\hat{X} = A_{\mathrm{re}}^{(k)}$, taking $\varepsilon^2 = 8C^2 kd$ according to Lemma 6. Condition (b) of the corollary holds if

$$32 C^2 (1+\kappa)^2 kd = 4(1+\kappa)^2 \varepsilon^2 < n_{1t} \delta_t^2(P) = \pi_{1t} \min_{s:\, s \neq t} \sum_{\ell=1}^{k_2} \pi_{2\ell} (\Psi_{s\ell} - \Psi_{t\ell})^2$$

for all $t \in [k_1]$, which is satisfied under assumption (37). Corollary 1, and specifically (26) gives the desired bound on misclassification rate $\leq 4(1+\kappa)^2 \varepsilon^2 / (n_1 \delta^2(P))$. ∎

Proof [Proof of Theorem 6] Recall that $A_{\mathrm{re}}^{(k)} = \hat{Z}_1 \hat{\Sigma} \hat{Z}_2^T$ is the k-truncated SVD of A_{re}. Let $X^{(1)} = \hat{Z}_1 \hat{\Sigma}$ and $X^{(2)} = A_{\mathrm{re}}^{(k)}$, and let $(x_i^{(1)})^T$ and $(x_i^{(2)})^T$ be their ith rows, respectively. Then,

$$\|x_2^{(i)} - x_2^{(j)}\| = \|\hat{Z}_2 (x_1^{(i)} - x_1^{(j)})\|_2 = \|x_1^{(i)} - x_1^{(j)}\|_2, \quad \forall i \neq j,$$

using $\hat{Z}_2 \in \mathbb{O}^{n_2 \times k}$ and (1). Isometry-invariance of $\mathscr{K}$ implies $\overline{\mathrm{Mis}}(\mathscr{K}(\hat{Z}_1 \hat{\Sigma}), \mathscr{K}(A_{\mathrm{re}}^{(k)})) = 0$. Since $\overline{\mathrm{Mis}}$ is a pseudo-metric on k-means matrices, using the triangle inequality, we get

$$\overline{\mathrm{Mis}}\left(\mathscr{K}(\hat{Z}_1 \hat{\Sigma}), P\right) \leq \overline{\mathrm{Mis}}\left(\mathscr{K}(\hat{Z}_1 \hat{\Sigma}), \mathscr{K}(A_{\mathrm{re}}^{(k)})\right) + \overline{\mathrm{Mis}}\left(\mathscr{K}(A_{\mathrm{re}}^{(k)}), P\right)$$

$$= \overline{\mathrm{Mis}}\left(\mathscr{K}(A_{\mathrm{re}}^{(k)}), P\right).$$

(In fact, using the triangle inequality in the other direction, we conclude that the two sides are equal.) The result now follows from Theorem 5. ∎

Proof [Proof of Corollary 4] We have

$$\|\Lambda_{s*} - \Lambda_{t*}\|^2 = \sum_{\ell=1}^{k_2} n_{2\ell}^2 (B_{s\ell} - B_{t\ell})^2 = \sum_{\ell=1}^{k_2} \frac{n_{2\ell}^2}{n_1 n_2} (\Psi_{s\ell} - \Psi_{t\ell})^2 = \alpha \sum_{\ell=1}^{k_2} \pi_{2\ell}^2 (\Psi_{s\ell} - \Psi_{t\ell})^2$$

where we have used $\alpha := n_2/n_1$. Recall from (42) that $\Lambda_\wedge^2 := \min_{t \neq s} \|\Lambda_{s*} - \Lambda_{t*}\|^2$. It follows that $\alpha^{-1} \Lambda_\wedge^2 \leq \Psi_{1,\wedge}^2$, using $\pi_{2,\ell} \leq 1$. We also have $\widetilde{\Psi}_{1,\wedge}^2 \geq \pi_{1,\wedge} \Psi_{1,\wedge}^2 \geq \pi_{1,\wedge} \alpha^{-1} \Lambda_\wedge^2$. Recalling the definition of a from (8), and using $\Psi_{s\ell} = (\sqrt{n_1 n_2}/n_{2\ell}) \Lambda_{s\ell}$, we have

$$d = \sqrt{\frac{n_2}{n_1}} \|\Psi\|_\infty \leq \sqrt{\frac{n_2}{n_1}} \frac{\sqrt{n_1 n_2}}{n_{2,\wedge}} \|\Lambda\|_\infty = \frac{1}{\pi_{2,\wedge}} \|\Lambda\|_\infty \leq \beta_2 k_2 \|\Lambda\|_\infty.$$

Hence, $kd\Psi_{1,\wedge}^{-2} \leq k\,\beta_2 k_2 \|\Lambda\|_\infty (\alpha^{-1}\Lambda_\wedge^2)^{-1} = \beta_2 kk_2\alpha\|\Lambda\|_\infty\Lambda_\wedge^{-2}$ which is the desired bound. We also note that

$$kd\widetilde{\Psi}_{1,\wedge}^{-2} \leq kd\pi_{1,\wedge}^{-1}\,\Psi_{1,\wedge}^{-2} \leq (\beta_1 k_1)\,kd\Psi_{1,\wedge}^{-2}$$

which combined with the previous bound shows that the required condition (43) in the statement is enough to satisfy (37). ∎

A.5. Proof of Theorem 7

As in the proof of Theorem 4, we only need to verify conditions of Corollary 1, with $X^* = \bar{Z}_1 U_\psi Q \in \mathbb{R}^{n_1 \times k}$, so that κ-approximate k-means operator $\mathcal{P}_\kappa$ satisfies bound (10) of the k-means step. Recall that $\bar{Z}_1 \in \mathbb{O}^{n_1 \times k_1}$, $U_\psi \in \mathbb{O}^{k_1 \times k}$ and $Q \in \mathbb{O}^{k \times k}$. Clearly, X^* has exactly k_1 distinct rows, $u_s^T Q$, $s \in [k_1]$. Furthermore,

$$\delta^2 := \delta_\wedge^2(\bar{Z}_1 U_\psi Q) = \delta_\wedge^2(\bar{Z}_1 U_\psi) = \min_{(t,s):\,t\neq s} \|n_{1t}^{-1/2}u_t - n_{1s}^{-1/2}u_s\|_2^2$$

where the second equality is by (1). Letting $\|\cdot\| = \|\cdot\|_2$, and $x = \pi_{1t}^{-1/2}$ and $y = \pi_{1s}^{-1/2}$ for simplicity, and $z = (x, -y) \in \mathbb{R}^2$, we have

$$
\begin{aligned}
n_1\|n_{1t}^{-1/2}u_t - n_{1s}^{-1/2}u_s\|^2 &= \|\pi_{1t}^{-1/2}u_t - \pi_{1s}^{-1/2}u_s\|^2 \\
&= \|\begin{pmatrix} u_t & u_s \end{pmatrix}\begin{pmatrix} x \\ -y \end{pmatrix}\|^2 \\
&= z^T(U_\psi U_\psi^T)_{\mathcal{I}}\,z, \qquad\qquad \text{(where } \mathcal{I} = \{t,s\}\text{),} \\
&= \|z\|^2 - z^T(I_{k_1} - U_\psi U_\psi^T)_{\mathcal{I}}\,z^T.
\end{aligned}
$$

Since $I_{k_1} - U_\psi U_\psi^T$ is positive semidefinite, using (44), we have

$$0 \leq z^T(I_{k_1} - U_\psi U_\psi^T)_{\mathcal{I}}\,z^T \leq (1-\rho_1)\|z\|^2,$$

hence, $n_1\|n_{1t}^{-1/2}u_t - n_{1s}^{-1/2}u_s\|^2 \geq \rho_1\|z\|^2 = \rho_1(\pi_{1t}^{-1} + \pi_{1s}^{-1})$. It follows by (A1),

$$(n_1\delta^2)^{-1} \leq \frac{1}{\rho_1}\max_{t\neq s}\,(\pi_{1t}^{-1} + \pi_{1s}^{-1})^{-1} \leq \frac{1}{\rho_1}\Big(\frac{\beta_1}{2k_1}\Big).$$

We also have

$$n_{1t}\,\delta_t^2 = n_{1t}\min_{s:\,s\neq t}\|n_{1t}^{-1/2}u_t - n_{1s}^{-1/2}u_s\|_2^2 \;\geq\; \rho_1\big(1 + \frac{\pi_{1t}}{\pi_{1s}}\big) \;\geq\; \rho_1.$$

We obtain, with $\varepsilon^2 = C_2^2 kd/\sigma_k^2$ and $O = U_\psi Q$,

$$\overline{\mathrm{Mis}}(\mathcal{K}(\hat{Z}_1), \bar{Z}_1) = \overline{\mathrm{Mis}}(\mathcal{K}(\hat{Z}_1), \bar{Z}_1 O) \;\lesssim\; \frac{\varepsilon^2}{n_1\delta^2} \;\lesssim\; \frac{\beta_1}{\rho_1}\frac{k}{k_1}\frac{d}{\sigma_k^2}$$

as long as $4(1+\kappa)^2(n_{1t}\,\delta_t^2)^{-1}\varepsilon^2 = 4C_2^2(1+\kappa)^2\,\rho_1^{-1}\,kd/\sigma_k^2 < 1$, to guarantee that condition (b) of Corollary 1 holds, so that $\mathcal{P}_\kappa$ satisfies (10) with $c = 4(1+\kappa)^2$ as in (26). The proof is complete.

A.6. Proof of Theorem 8

The following concentration result for sub-Gaussian random matrices is well-known (Vershynin, 2018):

Lemma 7 *Let $A \in \mathbb{R}^{n_1 \times n_2}$ have independent sub-Gaussian entries, and $\sigma = \max_{i,j} \|A_{ij} - \mathbb{E}A_{ij}\|_{\psi_2}$. Then for any $t > 0$, we have*

$$\|A - \mathbb{E}A\|_{op} \le C\,\sigma(\sqrt{n_1} + \sqrt{n_2} + t), \tag{67}$$

with probability at least $1 - 2\exp(-t^2)$.

Let $t = \sqrt{n_1} + \sqrt{n_2}$ in Lemma 7, then $\|A - \mathbb{E}A\|_{op} \le 2C\sigma(\sqrt{n_1} + \sqrt{n_2})$ with probability at least $1 - 2e^{-(n_1+n_2)}$. We thus have the concentration bound (8) with $\sqrt{d} = 2\sigma(\sqrt{n_1} + \sqrt{n_2})$. The proof of Theorem 5 goes through and the result follows by replacing d with the upper bound $d \le 4\sigma^2(n_1 + n_2)$, and replacing $\Psi_{1,\wedge}^2$ and $\widetilde{\Psi}_{1,\wedge}^2$ with $n_1 n_2 B_{1,\wedge}^2$ and $n_1 n_2 \widetilde{B}_{1,\wedge}^2$, respectively.

A.7. Proofs of Section 6.4

Proof [Proof of Proposition 2] We have with the given probability,

$$
\begin{aligned}
\|A_{\mathrm{re}} - \widetilde{P}\|_{\mathrm{op}} &\le \|A_{\mathrm{re}} - P\|_{\mathrm{op}} + \|P - \widetilde{P}\|_{\mathrm{op}} \\
&\le C\sqrt{d} + \frac{1}{\sqrt{n_1 n_2}} \|P^0 - \widetilde{P}^0\|_{\mathrm{op}} \\
&\le C\sqrt{d} + \frac{1}{\sqrt{n_1 n_2}} \|P^0 - \widetilde{P}^0\|_F \le (C_0 + C)\sqrt{d}
\end{aligned}
$$

where we have used Theorem 2 to bound the first term using $d' = d = \sqrt{n_2/n_1}\|P^0\|_\infty = n_2\|P\|_\infty$. The last inequality uses assumption (52). The proof is complete. $\blacksquare$

Proof [Proof of Proposition 3]
Letting $f(x, y) := [\rho_0(x, y) - \widetilde{\rho}_0(x, y)]^2$,

$$\|P^0 - \widetilde{P}^0\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n f(X_i, Y_j) = \sum_{j \in \mathbb{Z}_n} \sum_{i \in \mathbb{Z}_n} f(X_i, Y_{i+j}),$$

where $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ is $\{1, \ldots, n\}$ viewed as a cyclic group of order n. We have

$$\mathbb{E}[f(X_1, Y_1)^2] = \|f\|_{L^2}^2 = \|\rho_0 - \widetilde{\rho}_0\|_{L^4}^4 = o(d^2)$$

and $\mathbb{E}[f(X_1, Y_1)] = \|f\|_{L^1} \le \|f\|_{L^2} = o(d)$. Note that $\{f(X_i, Y_j) : i, j \in \mathbb{Z}_n\}$ are not independent, however, for each $j \in \mathbb{Z}_n$, $\{f(X_i, Y_{i+j}) : i \in \mathbb{Z}_n\}$ are i.i.d.. Let $Z_j := \sum_{i \in \mathbb{Z}_n} f(X_i, Y_{i+j})$. By independence,

$$\mathrm{var}(Z_j) = n\,\mathrm{var}(f(X_1, Y_1)) \le n\|f\|_{L^2}^2.$$

Since $\mathbb{E}Z_j = n\|f\|_{L^1}$, by the Chebyshev inequality, for any $j \in \mathbb{Z}_n$,

$$\mathbb{P}(Z_j \geq nd) = \mathbb{P}(Z_j - \mathbb{E}Z_j \geq nd - n\|f\|_{L^1})$$
$$\leq \frac{n\|f\|_{L^2}^2}{(na - n\|f\|_{L^1})^2} = \frac{1}{n}\frac{o(d^2)}{(d - o(d))^2} = o\left(\frac{1}{n}\right).$$

It follows that $\mathbb{P}\left(\|\!\|P^0 - \widetilde{P}^0\|\!\|_F^2 \geq n^2 d\right) \leq \sum_{j \in \mathbb{Z}_n} \mathbb{P}(Z_j \geq nd) \leq n\, o(n^{-1}) = o(1)$. ∎

Appendix B. Alternative algorithm for the k-means step

In this appendix, we present a simple general algorithm that can be used in the k-means step, replacing the κ-approximate k-means solver used throughout the text. The algorithm is based on the ideas in Gao et al. (2017) and Yun and Proutiere (2014b), and the version that we present here acheives the misclassification bound $\varepsilon^2/(n\delta^2)$ needed in Step 3 of the analysis (Section 3.1) without necessarily optimizing the k-means objective function. We present the results using the terminology of the k-means matrices (with rows in $\mathbb{R}^d$) introduced in Section 3.3, although the algorithm and the resulting bound work for data points in any metric space.

Let $X \in \mathbb{M}_{n,m}^k$ be a k-means matrix and let us denote its centers, i.e. distinct rows, as $\{q_r(X), r \in [k]\}$. As in Definition 2, we write $\delta_r(X)$ and $n_r(X)$ for the rth cluster center separation and size, respectively, and $\delta_\wedge(X) = \min_r \delta_r(X)$ and $n_\wedge = \min_r n_r(X)$. Assume that we have an estimate $\hat{X} \in \mathbb{R}^{n \times d}$ of X, and let us write $d(i,j) := d(\hat{x}_i, \hat{x}_j)$, $i,j \in [n]$ for the pairwise distances between the rows of $\hat{X}$.

Algorithm 6 which is a variant of the one presented in Gao et al. (2017), takes these pairwise distances and outputs cluster estimates $\widehat{\mathcal{C}}_1, \ldots, \widehat{\mathcal{C}}_k \subset [n]$, after k recursive passes through the data. A somewhat more sophisticated version of this algorithm appears in Yun and Proutiere (2014b), where one also repeats the process for $i = 1, \ldots, \log n$ and radii $R_i = iR_1$ in an outer loop, producing clusters $\widehat{\mathcal{C}}_r^{(i)}, r \in [k]$; one then picks, among these $\log n$ possible clusterings, the one that minimizes the k-means objective. The variant in Yun and Proutiere (2014b) also leaves no unlabeled nodes by assigning the unlabeled to the cluster whose estimated center is closest. In the rest of this section, we will focus on the simple version presented in Algorithm 6 as this is enough to establish our desired bound. The following theorem provides the necessary guarantee:

Theorem 9 *Consider the cluster model above and let $n_r = n_r(X)$, $n_\wedge = n_\wedge(X)$ and $\delta_\wedge = \delta_\wedge(X)$. Assume that we have approximate data $\hat{x}_1, \ldots, \hat{x}_n$ such that $\sum_{i=1}^n d(x_i, \hat{x}_i)^2 \leq \varepsilon^2$. In addition, assume that for some $\gamma \in (0,1)$ and $\beta \geq 1$:*

(i) $n_r \leq \beta n_\wedge$ for all $r \in [k]$ (Clusters are β-balanced.),

(ii) $\dfrac{2\varepsilon}{\sqrt{\gamma n_\wedge}} < \dfrac{\delta_\wedge}{3}$ (ε^2 small enough compared to $n_\wedge \delta_\wedge^2$.),

(iii) $\xi\beta + \gamma < 1 - \gamma$ where $\xi := \gamma/(1-\gamma)$. (Gamma small enough relative to β.)

Algorithm 6 k-means replacement

Input: Pairwise distance $d(i, j)$, $i, j \in [n]$ and radius ρ.

1: $S \leftarrow [n]$
2: **for** $r = 1, \ldots, k$ **do**
3:     For every $i \in S$, let $B_d(i; \rho) := \{j \in S : d(i, j) \leq \rho\}$.
4:     Pick $i_0 \in S$ that maximizes $i \mapsto |B_d(i; \rho)|$.
5:     Let $\widehat{\mathcal{C}}_r = B_d(i_0; \rho)$.
6:     $S \leftarrow S \setminus \widehat{\mathcal{C}}_r$.
7: **end for**

Output: Return clusters $\widehat{\mathcal{C}}_r : 1, \ldots, k$ and output remaining S as unlabeled.

Let $M_n(\rho)$ be the (average) misclassification rate of Algorithm 6 with input radius ρ. Then,

$$M_n(\rho) \leq \frac{8\varepsilon^2}{n\rho^2}, \quad \forall \rho \in \Big[\frac{2\varepsilon}{\sqrt{\gamma n_\wedge}}, \frac{\delta_\wedge}{3} \Big).$$

Applying the algorithm with $\rho \geq \alpha\delta_\wedge$ for $\alpha < 1/3$ we obtain the misclassication bound $c_\alpha \varepsilon^2 /(n\delta_\wedge^2)$ where $c_\alpha = 8/\alpha^2$. Thus, Algorithm 6 with a proper choice of the radius ρ satisfies the desired bound (10) of the k-means step.

Proof [Proof of Theorem 9] The proof follows the argument in Gao et al. (2017). As in the proof of Proposition 1, let $\mathcal{C}_r$ denote the rth cluster of X, having center $q_r = q_r(X)$. We have $|\mathcal{C}_r| = n_r$. Let x_i and $\hat{x}_i$ be the ith row of X and $\hat{X}$, respectively, and let

$$T_r := \{i \in \mathcal{C}_r : d(\hat{x}_i, q_r) < \rho/2\} = \{i \in \mathcal{C}_r : d(\hat{x}_i, x_i) < \rho/2\}$$

using $x_i = q_r$ for all $i \in \mathcal{C}_r$ which holds by definition. $\{T_r\}$ are disjoint and clearly $T_r \subset \mathcal{C}_r$. Let $T := \biguplus_r T_r$, a disjoint union, and $T^c = [n] \setminus T$. We have

$$|T^c|\rho^2/4 \ \leq \ \sum_{i \in T^c} d(\hat{x}_i, x_i)^2 \leq \varepsilon^2 \ \implies \ |T^c| \leq 4\varepsilon^2/\rho^2. \tag{68}$$

As a consequence of assumption (ii) and our choice of ρ, we have $4\varepsilon^2/(n_\wedge \rho^2) \leq \gamma$, hence

$$|T_r| = |\mathcal{C}_r| - |\mathcal{C}_r \setminus T_r| \ \geq \ |\mathcal{C}_r| - |T^c| \ \geq \ n_\wedge \Big(1 - \frac{4\varepsilon^2}{n_\wedge \rho^2}\Big) \geq n_\wedge(1 - \gamma) \tag{69}$$

for all $r \in [k]$. On the other hand, $|T^c| \leq \gamma n_\wedge$. In particular, combining the two estimates

$$|T^c| \leq \xi \, |T_r|, \quad \forall r \in [k] \tag{70}$$

where $\xi = \gamma/(1-\gamma)$. These size estimates will be used frequently in the course of the proof.

Recall that $d(i, j) := d(\hat{x}_i, \hat{x}_j)$, $i, j \in [n]$, the collection of pairwise distances between the data points $\hat{x}_1, \ldots, \hat{x}_n$. Thus, with some abuse of notation, $(i, j) \mapsto d(i, j)$ defines a pseudo-metric on $[n]$ (and a proper metric if $\{\hat{x}_i\}$ are distinct). For any two subsets $A, B \subset [n]$ we write $d(A, B) = \inf\{d(i, j) : i \in A, j \in B\}$. For any $i \in [n]$, let $d(i, A) = d(\{i\}, A)$.

We say that node i_0 is near T if $d(i_0, T) \leq \rho$, i.e., i_0 belongs to the ρ-enlargement of T. Similarly, we say that i_0 is near T_r if $d(i_0, T_r) \leq \rho$ and far from T_r otherwise. Note that i_0

can be near at most one of $T_r, r \in [k]$. This is since $d(T_r, T_\ell) \geq \delta_\wedge - \rho$ for $r \neq \ell$, and we are assuming $\delta_\wedge > 3\rho$. In fact, i_0 *is near T iff i_0 is near exactly one of $T_r, r \in [k]$.*

To understand Algorithm 6, let us assume that we are at some iteration of the algorithm and we are picking the center i_0 and the corresponding cluster $\widehat{\mathcal{C}} := \{j : d(j, i_0) \leq \rho\}$. One of the following happens:

(a) We pick the new center $i_0 \in T_r$ for some r, in which case $\widehat{\mathcal{C}}$ will include the entire T_r, none of $T_\ell, \ell \neq r$, and perhaps some of T^c. That is, $\widehat{\mathcal{C}} \supset T_r$ and $\widehat{\mathcal{C}} \cap T_\ell = \emptyset$ for $\ell \neq r$.

(b) We pick i_0 *near* T_r for some r. In this case, $|\widehat{\mathcal{C}}| \geq |T_r|$, otherwise any member of T_r would have created a bigger cluster by part (a) above. Now, $\widehat{\mathcal{C}}$ cannot contain any of $T_\ell, \ell \neq r$, because i_0 is far from those if it is near T_r. Hence, $\widehat{\mathcal{C}} \subset T_r \cup T^c$. Since $|T^c| \leq \xi|T_r|$ by (70), and $|\widehat{\mathcal{C}}| \geq |T_r|$, we have $|\widehat{\mathcal{C}} \cap T_r| \geq (1-\xi)|T_r|$. That is, $\widehat{\mathcal{C}}$ contains a large fraction of T_r.

If either of the two cases above happen, we say that T_r is depleted, otherwise it is intact. If T_r is depleted, it will not be revisited in future iterations, as long as other intact $T_\ell, \ell \neq r$ exist. To see this, first note that $|T_r \cap \widehat{\mathcal{C}}^c| \leq \xi|T_r| \leq \xi\beta n_\wedge$, using assumption (i). Taking i_0 on or near T_r in a future iteration will give us a cluster of size at most $(\xi\beta + \gamma)n_\wedge < (1-\gamma)n_\wedge$ (by assumption (iii)) which is less that $|T_\ell|$ for an intact cluster.

To simplify notation, if either of (a) and (b) happen, i.e., we pick cluster center i_0 near T_r for some r, we name the corresponding cluster $\widehat{\mathcal{C}}_r$. This is to avoid carrying around a permutation of cluster labels different than the original one, and is valid since each T_r is visited at most once by the above argument. (In fact, each is visited exactly once, as we argue below.) That last possibility is

(c) We pick i_0 *far* from any T_r, that is $d(i_0, T) > \rho$. This gives $\widehat{\mathcal{C}} \subset T^c$, hence $|\widehat{\mathcal{C}}| \leq |T^c| \leq \gamma n_\wedge < (1-\gamma)n_\wedge \leq |T_\ell|$ for any intact T_ℓ. Thus as along as there are intact T_ℓ, this case does not happen.

The above argument gives the following picture of the evolution of the algorithm: At each step $t = 1, \ldots, k$, we pick i_0 near T_r for some previously unvisited r, making it depleted, creating estimated cluster $\widehat{\mathcal{C}}_r$ and proceeding to the next iteration. After the k-th iteration all $T_\ell, \ell \in [k]$ will be depleted. We have $|\widehat{\mathcal{C}}_r| \geq |T_r|$, and $\widehat{\mathcal{C}}_r \subset T_r \cap T^c$ for all $r \in [k]$.

By construction $\{\widehat{\mathcal{C}}_\ell\}$ are disjoint. Let $\widehat{\mathcal{C}} := \biguplus_{\ell \in [k]} \widehat{\mathcal{C}}_\ell$, and note that $|\widehat{\mathcal{C}}| \geq |T|$ hence $|\widehat{\mathcal{C}}^c| \leq |T^c|$. Since $\widehat{\mathcal{C}}_\ell \cap T_r = \emptyset$ for $\ell \neq r$, we have $T_r \subset \bigcap_{\ell \neq r} \widehat{\mathcal{C}}_\ell^c$, hence $T_r \cap \widehat{\mathcal{C}}_r^c \subset \widehat{\mathcal{C}}^c$. All the misclassified or unclassified nodes produced by the algorithm are contained in $[\bigcup_r (T_r \cap \widehat{\mathcal{C}}_r^c)] \cup T^c$ which itself is contained in $\widehat{\mathcal{C}}^c \cup T^c$. Hence, the misclassification rate is bounded above by

$$\frac{1}{n}|\widehat{\mathcal{C}}^c \cup T^c| \leq \frac{1}{n}(|\widehat{\mathcal{C}}^c| + |T^c|) \leq \frac{2}{n}|T^c| \leq \frac{8\varepsilon^2}{n\rho^2}.$$

where we have used (68). The proof is complete. ∎

Remark 13 *The last part of the argument can be made more transparent as follows: Each $\widehat{\mathcal{C}}_r$ consists of two disjoint part, $\widehat{\mathcal{C}}_r \cap T_r$ and $\widehat{\mathcal{C}}_r \cap T^c$. We have $|\widehat{\mathcal{C}}_r \cap T^c| \geq |T_r \setminus \widehat{\mathcal{C}}_r|$ (equivalent to $|\widehat{\mathcal{C}}_r| \geq |T_r|$). Then,*

$$\sum_k |T_r \setminus \widehat{\mathcal{C}}_r| \leq \sum_r |\widehat{\mathcal{C}}_r \cap T^c| = |\widehat{\mathcal{C}} \cap T^c| \leq |T^c|$$

and total misclassifications are bounded by $\sum_k |T_r \setminus \widehat{\mathcal{C}}_r| + |T^c|$.

Appendix C. Auxiliary lemmas

Lemma 8 *Let $Z, Y \in \mathbb{O}^{n \times k}$ and let $\Pi_Z = ZZ^T$ and $\Pi_Y = YY^T$ be the corresponding projection operators. We have*

$$\min_{Q \in \mathbb{O}^{k \times k}} \|Z - YQ\|_F \leq \|\Pi_Z - \Pi_Y\|_F.$$

Proof We first note that $\|\Pi_Z\|_F^2 = \operatorname{tr}(\Pi_Z^2) = \operatorname{tr}(\Pi_Z) = k$ (since projections are idempotent), and $\|Z\|_F^2 = \operatorname{tr}(Z^T Z) = \operatorname{tr}(\Pi_Z) = k$. Let $Z^T Y = U\Sigma V^T$ be the SVD of $Z^T Y$ where $U, V \in \mathbb{O}^{k \times k}$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k) \succeq 0$. Then, using the change of variable $O = V^T Q U$,

$$\frac{1}{2} \min_Q \|Z - YQ\|_F^2 = k - \max_Q \operatorname{tr}(Z^T Y Q)$$
$$= k - \max_{O \in \mathbb{O}^{k \times k}} \operatorname{tr}(\Sigma O) = k - \|\Sigma\|_*,$$

where $\|\Sigma\|_* = \sum_i \sigma_i$ is the nuclear norm of Σ. To see the last equality, we note that since O is orthogonal, we have $|O_{ii}| \leq 1$ for all i, hence $\max_O \operatorname{tr}(\Sigma O) \leq \max_{\forall i, \, |O_{ii}| \leq 1} \sum_i \sigma_i O_{ii} = \sum_i \sigma_i$ by the duality of ℓ_1 and ℓ_∞ norms and $\sigma_i \geq 0$. The equality is achieved by $O = I_k$. On the other hand

$$\frac{1}{2}\|\Pi_Z - \Pi_Y\|_F^2 = k - \operatorname{tr}(\Pi_Z \Pi_Y) = k - \|Z^T Y\|_F^2 = k - \|\Sigma\|_F^2.$$

Since $\|\Sigma\|_{\mathrm{op}} = \|Z^T Y\|_{\mathrm{op}} = \|Z\|_{\mathrm{op}} \|Y\|_{\mathrm{op}} \leq 1$, we have $\sigma_i \leq 1$ for all i. It follows that $\|\Sigma\|_F^2 \leq \|\Sigma\|_*$ completing the proof. ∎

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962, 2011.

Afonso S Bandeira, Ramon Van Handel, et al. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, pages pnas–0907096106, 2009.

Norbert Binkiewicz, Joshua T. Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. pages 1–48, nov 2014. URL `http://arxiv.org/abs/1411.2158`.

Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

Ravi B Boppana. Eigenvalues and graph bisection: An average-case analysis. In *Foundations of Computer Science, 1987., 28th Annual Symposium on*, pages 280–285. IEEE, 1987.

Joshua Cape, Minh Tang, and Carey E Priebe. On spectral embedding performance and elucidating network structure in stochastic block model graphs. *arXiv preprint arXiv:1808.04855*, 2018.

Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pages 35–1, 2012.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *The Journal of Machine Learning Research*, 17(1):882–938, 2016.

Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423, 2015.

Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

Anirban Dasgupta, John Hopcroft, Ravi Kannan, and Pradipta Mitra. Spectral clustering by recursive partitioning. *Algorithms–ESA 2006*, pages 256–267, 2006.

Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. 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, 2013.

Chao Gao, Yu Lu, Harrison H Zhou, et al. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024, 2017.

Chao Gao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.

John A Hartigan. Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129, 1972.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

A Joseph and B Yu. Impact of regularization on Spectral Clustering. *arXiv preprint arXiv:1312.1733*, 2013. URL http://arxiv.org/abs/1312.1733.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83(1):016107, 2011.

Olga Klopp, Alexandre B Tsybakov, Nicolas Verzelen, et al. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences of the United States of America*, 110(52):20935–40, dec 2013. ISSN 1091-6490. doi: 10.1073/pnas.1312486110. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3876200{&}tool=pmcentrez{&}rendertype=abstract.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+/spl 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.

45

C. M. Le, E. Levina, and R. Vershynin. Sparse random graphs: regularization and concentration of the Laplacian. *arXiv preprint arXiv:1502.03049*, feb 2015. doi: 10.1088/0264-9381/32/11/115009. URL http://arxiv.org/abs/1502.0304http://arxiv.org/abs/1502.00304http://arxiv.org/abs/1502.03049.

Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 2017.

Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Vince Lyzinski, Daniel L Sussman, Minh Tang, Avanti Athreya, Carey E Priebe, et al. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, 2014.

F. McSherry. Spectral partitioning of random graphs. In *Proceedings 2001 IEEE International Conference on Cluster Computing*, pages 529—-537. IEEE Comput. Soc, 2001. ISBN 0-7695-1116-3. doi: 10.1109/SFCS.2001.959929. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=959929.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

Sofia C Olhede and Patrick J Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

Karl Rohe and Bin Yu. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *stat*, 1050:10, 2012.

Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Bo Söderberg. General formalism for inhomogeneous random graphs. *Physical review E*, 66 (6):066121, 2002.

Liangjun Su, Wuyi Wang, and Yichong Zhang. Strong consistency of spectral clustering for stochastic block models. *arXiv preprint arXiv:1710.06191*, 2017.

D.-C. Tomozei and L. Massoulié. Distributed user profiling via spectral methods. In *ACM SIGMETRICS Performance Evaluation Review*, pages 383—-384, 2010. URL http://www.i-journals.org/ssy/viewarticle.php?id=36http://arxiv.org/abs/1109.3318http://dl.acm.org/citation.cfm?id=1811098.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

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

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. *arXiv preprint arXiv:1709.03183*, 2017.

Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014a.

Se-Young Yun and Alexandre Proutiere. Community detection via random and adaptive sampling. In *Conference on Learning Theory*, pages 138–175, 2014b.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40 (4):2266–2292, 2012.

Zhixin Zhou and Arash A. Amini. Optimal bipartite network clustering. *Preprint*, 2018.

Zhixin Zhou and Arash A. Amini *GitHub repository: Bipartite Spectral Clustering*. URL https://github.com/zhixin0825/bipartite-spectral-clustering, 2018.