Spectral label recovery for Stochastic block models

December 10, 2015

Abstract: In these short notes, we provide an account of recent results on latent class recovery for Stochastic Block Models (SBM), focusing on spectral clustering methods.

Disclaimer: These notes complement the last lecture of a short course on Stochastic block models given at ENSAE[1]. The reader is assumed to be familiar with the notion of SBM (see e.g. [25] for an introduction).

Notations

- $\boldsymbol{A}$ adjacency matrix of the graph $\mathcal{G}$.
- $\boldsymbol{B}$ probability matrix of size $K \times K$
- $Z = (Z_i)_{i=1}^n$ vector of (unknown) labels
- $\boldsymbol{M}_{n,K}$ collection of $n \times K$ matrices where each row has exactly one 1 and $(K-1)$ zero.
- $\boldsymbol{\Theta} \in \boldsymbol{M}_{n,K}$ membership matrix defined by $\boldsymbol{\Theta}_{ik} = 1$ if and only if $Z_i = k$.
- $G_k := \{i : Z_i = k\}$ collection of nodes with label k.
- $n_k := |G_k|$
- $n_{min} := \min_{k=1,\dots,K}(n_k)$
- $\boldsymbol{P} := \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{\Theta}^*$. All the off-diagonal entries $\boldsymbol{P}_{i,j}$ of $\boldsymbol{P}$ are equal to $\mathbb{E}[\boldsymbol{A}_{i,j}]$.
- Σ_K permutation group of K elements.
- $\mathcal{L}(.)$ symmetric Laplacian. Given a $n \times n$ symmetric $\boldsymbol{A}$ with nonnegative entries, $\mathcal{L}(\boldsymbol{A})$ is defined by $\mathcal{L}(\boldsymbol{A}) := \boldsymbol{D}_{\boldsymbol{A}}^{-1/2}\boldsymbol{A}\boldsymbol{D}_{\boldsymbol{A}}^{-1/2}$ where the diagonal matrix $\boldsymbol{D}_{\boldsymbol{A}}$ satisfies $(\boldsymbol{D}_{\boldsymbol{A}})_{i,i} = \sum_{j=1}^n \boldsymbol{A}_{i,j}$.
- $\lambda_{\min}(.)$ and $\lambda_{\max}()$ smallest and the largest eigenvalues of a matrix.

[1]Please send any feedback (comments, typos,...) to nicolas.verzelen@supagro.inra.fr

- **Norms**: $\|.\|_F$ stands for the Frobenius norm. Given a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ is the spectral norm, that is the largest singular value of $\boldsymbol{A}$. For any vector u, $\|u\|$ denotes the Euclidean norm.
- Given a matrix $\boldsymbol{X}$, $\boldsymbol{X}_{k*}$ denotes the k-th row of $\boldsymbol{X}$.

Contents

1 Introduction

Definition 1 (Stochastic block model with labeling Z)**.** *Let $\boldsymbol{B}$ be a $K \times K$ symmetric matrix with entries in $[0,1]$ and let $Z \in \{1,\dots,K\}^n$. The symmetric random matrix $\boldsymbol{A}$ is distributed according to a stochastic block model $\tilde{\mathbb{G}}(n,\boldsymbol{B},Z)$ with fixed labels Z if the diagonal entries of $\boldsymbol{A}$ zero and the upper diagonal entries of $\boldsymbol{A}$ satisfy $\mathbb{P}[\boldsymbol{A}_{i,j}=1] = \boldsymbol{B}_{Z_i,Z_j}$.*

Remark: In contrast to the definition of Stochastic block model used in the first lectures, the vector Z of labels is now seen as a fixed unknown parameter.

Objective: Given the adjacency matrix $\boldsymbol{A}$, our goal is to recover the vector Z of labels (up to possible permutations of $\{1,\dots,K\}$). The matrix $\boldsymbol{B}$ is assumed to be unknown, but the number K of classes is known. For any $k=1,\dots,K$, we denote $G_k := \{i : Z_i = k\}$ the collection of indices of label k. Also, $\boldsymbol{\Theta} \in \boldsymbol{M}_{n,K}$ is the membership matrix defined by $\boldsymbol{\Theta}_{ik} = 1$ if and only if $Z_i = k$. It is equivalent to estimate the label Z vector, the groups (G_k) or the membership matrix $\boldsymbol{\Theta}$ and we shall equally write $\widehat{Z}$, $(\widehat{G}_k)_{k=1}^K$, or $\widehat{\boldsymbol{\Theta}}$ for an estimation of the labels.

1.1 Weak, strong and perfect recovery

Given a predictor $\widehat{Z}$ of Z, we measure the quality of $\widehat{Z}$ by

$$l_*(\widehat{Z};Z) := \inf_{\sigma\in\Sigma_K} \sum_{k=1}^{K} \sum_{i\in G_k} \frac{\mathbf{1}_{\widehat{Z}_i \neq \sigma(Z_i)}}{n_i} \ , \tag{1}$$

where the infimum is taken over all permutations of $\{1,\dots,K\}$.

Although most results described in these notes are non-asymptotic, we shall sometimes interpret them in an asymptotic way where n goes to infinity while K and $\boldsymbol{B}$ possibly vary with n. We say that a reconstruction procedure achieves **weak recovery**[2] when

$$\underline{\lim}\, \mathbb{E}\left[\frac{l_*(\widehat{Z},Z)}{K-1}\right] < 0 \ . \tag{2}$$

Note that condition (2) means that $\widehat{Z}$ performs better than any random guess reconstruction method (a random guess method does not have access to the adjacency matrix $\boldsymbol{A}$ to predict Z).

A reconstruction procedure is said to achieve **strong recovery** when $l_*(\widehat{Z},Z) = o_P(1)$. Although this is not the focus of these notes, we say thats $\widehat{Z}$ achieves **perfect recovery**, when $l_*(\widehat{Z},Z) = 0$ with probability going to one.

1.2 Summary

There is a extensive literature on label recovery for stochastic block models. In these notes, we are interested in polynomial methods that provably work for (almost) arbitrary probability matrices $\boldsymbol{B}$. Section 2 is devoted to the analysis of various spectral clustering methods. Singular value thresholding methods are discussed in Section 3. In order to assess the optimality of these methods, a minimax lower bound is provided in Section 4. We leave out maximum likelihood methods [8, 10]

[2]The definition of weak recovery is slightly different from the one in the earlier lectures as we now use the loss $l(.,.)$. This choice of error measure is driven here by the proof techniques of Section 4.

and profile likelihood methods [37] as these estimators are not provably computed in polynomial time and the known reconstruction bounds do not improve over spectral methods. Also, we will not discuss convex relaxations of the maximum likelihood estimator (e.g. [5, 11, 17]) as their analysis is restricted to strongly assortative matrices $\boldsymbol{B}$ (the smallest diagonal value of B is larger than the largest off-diagonal value). Other approaches such as tensor product estimation [6] are briefly discussed in Section 5. In these notes, we focus on the general strategy for proving the reconstruction bounds.

2 Spectral clustering procedures

Throughout this section, it is assumed that $\boldsymbol{B}$ is full rank.

2.1 Eigenstructure of $\boldsymbol{P}$

Recall that $\boldsymbol{\Theta} \in \boldsymbol{M}_{n,K}$ is the membership matrix defined by $\boldsymbol{\Theta}_{ik} = 1$ if and only if $Z_i = k$. Then, the matrix $\boldsymbol{P} = \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{\Theta}^*$ satisfies $\boldsymbol{P}_{i,j} = \mathbb{E}[\boldsymbol{A}_{i,j}]$ for all $i \neq j$. In order to motivate the use of the first eigenvectors of $\boldsymbol{A}$, let us first characterize the eigenstructure of $\boldsymbol{P}$. Given a matrix $\boldsymbol{X}$, $\boldsymbol{X}_{k*}$ denotes its k-th row.

Lemma 1 ([22])**.** *Let $(Z, \boldsymbol{B})$ parametrize a SBM with K communities, where $\boldsymbol{B}$ is full rank. Let $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ be a spectral decomposition of $\boldsymbol{P} = \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{\Theta}^*$. Then $\boldsymbol{U} = \boldsymbol{\Theta}\boldsymbol{X}$ where $\boldsymbol{X} \in \mathbb{R}^{K\times K}$ and $\|\boldsymbol{X}_{k*} - \boldsymbol{X}_{l*}\| = \sqrt{n_k^{-1} + n_l^{-1}}$.*

Proof of Lemma 1. Let Δ be the diagonal matrix with diagonal elements $(\sqrt{n_1}, \ldots, \sqrt{n_K}$, then $\boldsymbol{P}$ decomposes as

$$\boldsymbol{P} = \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{\Theta}^* = \boldsymbol{\Theta}\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}B\boldsymbol{\Delta}(\boldsymbol{\Theta}\boldsymbol{\Delta}^{-1})^* \ .$$

Let $\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^* = \boldsymbol{\Delta}\boldsymbol{B}\boldsymbol{\Delta}$ be a spectral decomposition of $\boldsymbol{\Delta}\boldsymbol{B}\boldsymbol{\Delta}$. Upon defining $\boldsymbol{U} = \boldsymbol{\Theta}\boldsymbol{\Delta}^{-1}\boldsymbol{Z}$, we observe $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ is a spectral decomposition of $\boldsymbol{P}$. Since $\boldsymbol{Z}$ is an orthogonal matrix, the rows of the matrix $\boldsymbol{X} = \boldsymbol{\Delta}^{-1}Z$ are orthogonal with l_2 norm $\|\boldsymbol{X}\_k*\| = \sqrt{1/n_k}$. □

As a consequence, the $n \times K$ matrix $\boldsymbol{U}$ only contains K different rows, each row being identified with one of the K latent classes. It is therefore tempting to estimate the $n \times K$ matrix $\widehat{\boldsymbol{U}}$ of eigenvectors associated to the K largest (in absolute value) eigenvalues of the Adjacency matrix $\boldsymbol{A}$.

Rather than studying the eigenvectors of the adjacency matrix $\boldsymbol{A}$, it is customary to consider the Laplacian of $\boldsymbol{A}$. Here, we consider the symmetric Laplacian $\mathcal{L}(.)$ defined by $\mathcal{L}(\boldsymbol{A}) = \boldsymbol{D}_{\boldsymbol{A}}^{-1/2}\boldsymbol{A}\boldsymbol{D}_{\boldsymbol{A}}^{-1/2}$ where the diagonal matrix $\boldsymbol{D}_{\boldsymbol{A}}$ satisfies $(\boldsymbol{D}_{\boldsymbol{A}})_{i,i} = \sum_{j=1}^n \boldsymbol{A}_{i,j}$. As in Lemma 1, the membership matrix $\boldsymbol{\Theta}$ is characterized by the eigenstructure of $\mathcal{L}(\boldsymbol{P})$

Lemma 2 (**Eigenstructure of the Laplacian $\mathcal{L}(\boldsymbol{P})$**)**.** *Let $(Z, \boldsymbol{B})$ parametrize a SBM with K communities, where $\boldsymbol{B}$ is full rank. Let $\boldsymbol{U}_L\boldsymbol{D}_L\boldsymbol{U}_L^*$ be a spectral decomposition of $\mathcal{L}(\boldsymbol{P}) = \mathcal{L}(\boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{\Theta}^*)$. Then $\boldsymbol{U}_L = \boldsymbol{\Theta}\boldsymbol{X}$ where $\boldsymbol{X} \in \mathbb{R}^{K\times K}$ and $\|\boldsymbol{X}_{k_*} - \boldsymbol{X}_{l_*}\| = \sqrt{n_k^{-1} + n_l^{-1}}$.*

2.2 Algorithm

Let us first start with a simple example of two-class stochastic block model with $n_1 = n_2 = 30$ nodes and a probability matrix $\boldsymbol{B} = \begin{pmatrix} 0.7 & 0.4 \\ 0.4 & 0.7 \end{pmatrix}$. According to Lemma 1, the population matrix $\boldsymbol{P}$ contrains two non-zero eigenvalues and the $n \times 2$ matrix $\boldsymbol{U}$ of eigenvectors only contains two

distinct rows, one for each community. As depicted in Figure 1, the two largest eigenvalues of $\boldsymbol{A}$ are close to those of $\boldsymbol{P}$ and are well separated from the bulk distribution. Furthermore, the rows of the $n \times 2$ matrix $\widehat{\boldsymbol{U}}$ are close to those of $\boldsymbol{U}$ (Figure 2). Thus, rows $\widehat{\boldsymbol{U}}_{k,*}$ corresponding to nodes in the same community tend be close to each other. Any clustering method building groups according to the distance between rows of $\widehat{\boldsymbol{U}}$ should approximately recover the true communities.

Algorithm 1 below describes the general recipe for spectral community reconstruction. The procedure requires the specification of a function $\Psi : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ and a Distance-based clustering method *DistanceClustering* that groups the rows of $\widehat{\boldsymbol{U}}$ into K groups. When the graph is not too sparse, a simple choice is $\Psi = \mathrm{Id}$, in which case Algorithm 1 is called Adjacency-spectral clustering. Alternatively, the choice $\Psi(.) = \mathcal{L}(.)$ corresponds to the Laplacian spectral clustering[3].

Algorithm 1 Simple spectral algorithms

Require: $\boldsymbol{A}$, K, Ψ, *DistanceClustering*

1- Compute the $n \times K$ matrix $\widehat{\boldsymbol{U}}$ of the eigenvectors associated with the K largest eigenvalues of $\Psi(\boldsymbol{A})$ (in absolute value).

2- Run *DistanceClustering*$(\widehat{\boldsymbol{U}}, K)$.

Output: Partition $\widehat{G}_1, \ldots, \widehat{G}_K$ of the nodes.

If $\widehat{\boldsymbol{U}}$ was equal to $\boldsymbol{U}$ (as defined in Lemma 1) or $\boldsymbol{U}_L$ (as defined in Lemma 2), then one could easily recover the partition $(G_1, \ldots, G_K)$ of the nodes by assigning the same labels to rows of $\boldsymbol{U}$ that are identical. In practice, $\widehat{\boldsymbol{U}}$ is a noisy version of $\boldsymbol{U}$ and we use *DistanceClustering* to partition the rows of $\widehat{\boldsymbol{U}}$ in K classes $(\widehat{G}_1, \ldots, \widehat{G}_K)$ in such a way that rows of $\widehat{\boldsymbol{U}}$ that are close to each other tend to be in the same class.

In these notes, we only study the performances of the clustering procedure when *DistanceClustering* is an approximation of the K-means problem (see the next subsection for definitions). Indeed, approximate K-means solutions are simple to analyze and can be computed in polynomial time (with respect to n or K).

2.3 Approximate K-means algorithms

Given a matrix $V \in \mathbb{R}^{n\times p}$, the K-means minimization problem is defined by

$$(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{X}}) := \arg \min_{\boldsymbol{\Theta}\in \boldsymbol{M}_{n,K},\ X\in\mathbb{R}^{K\times p}} \|\boldsymbol{\Theta}\boldsymbol{X} - \boldsymbol{V}\|_F^2 \ . \tag{3}$$

The rows of $\widehat{\boldsymbol{X}}$ are called the centroids and $\widehat{\boldsymbol{\Theta}}$ is the membership matrix of the clustering. Denoting $\boldsymbol{V}_{i,*} \in \mathbb{R}^p$ the i-th row of $\boldsymbol{V}$, we observe that (3) is equivalent to finding a partition $(\widehat{G}_1, \ldots, \widehat{G}_K)$ of $\{1, \ldots, n\}$ and K centroids $\boldsymbol{X}_{1,*}, \ldots, \boldsymbol{X}_{K,*}$ in $\mathbb{R}^p$ in such a way that the sum of the square euclidean distance $\sum_{k=1}^K \sum_{i\in\widehat{G}_k} \|\boldsymbol{V}_{i,*} - \boldsymbol{X}_{k,*}\|^2$ is minimized.

However, the minimization problem in (3) is NP-hard. Iterative methods such as Lloyd's algorithm only converge to a local minimum. As customary in computational complexity, we therefore consider the simpler problem of finding an approximate solution of the problem. Fix $c > 1$. $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{X}})$ is said to be a c**-approximate solution of** K**-means** if

$$\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{X}} - \boldsymbol{V}\|_F^2 \le c \min_{\boldsymbol{\Theta}\in \boldsymbol{M}_{n,K},\ \boldsymbol{X}\in\mathbb{R}^{K\times p}} \|\boldsymbol{\Theta}\boldsymbol{X} - \boldsymbol{V}\|_F^2 \ .$$

[3] Actually, Algorithm 1 differs from the usual definition [33] of spectral clustering where the K largest eigenvalues of $\mathcal{L}(\boldsymbol{A})$ are considered (instead of the K largest in absolute values)

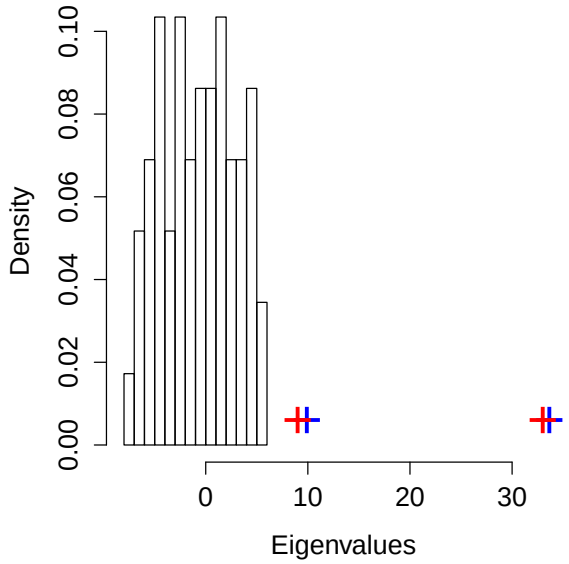

Figure 1: The two non-zero eigenvalues of the population matrix $\boldsymbol{P}$ are depicted in red, the two largest largest eigenvalues of the adjacency matrix $\boldsymbol{A}$ being in blue. The remaining eigenvalues of $\boldsymbol{A}$ are summarized by an histogram.

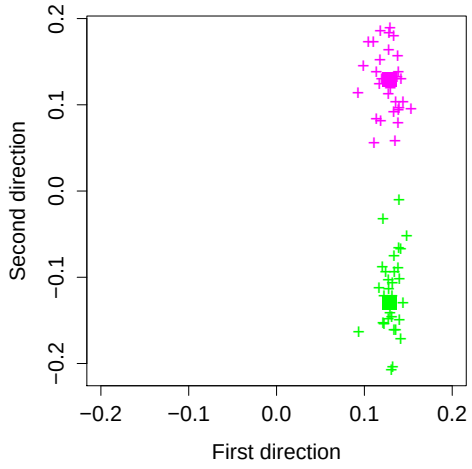

Figure 2: Coordinates of two first eigenvectors of $\boldsymbol{A}$. Nodes in the first (resp. second) community are depicted in green (resp. magenta). Green and Magenta squares correspond to the coordinates of two non zero eigenvectors $\boldsymbol{P}$ (these coordinates are equal for all nodes belonging to the same community).

For ϵ arbitrarily small, it is possible to compute $(9+\epsilon)$-approximate solutions in polynomial time [19] (with respect to n or K). Conversely, the problem of finding an $1+\epsilon$-approximate solutions with a small ϵ (say $\epsilon = 1$) has recently been proved to be hard [7].

Coming back to the spectral clustering problem, we consider an approximation of the K-means problem for $\widehat{\boldsymbol{U}} \in \mathbb{R}^{n\times K}$ (note that $p = K$ in our setting). In the following lemma, we characterize the clustering error of any approximate solution $\widehat{\boldsymbol{\Theta}}$ of the K-means problem in terms of the difference between $\widehat{\boldsymbol{U}}$ and its target $\boldsymbol{U}$.

Lemma 3 ([22]). *For any $\epsilon > 0$ and any two matrices $\widehat{\boldsymbol{U}}$, $\boldsymbol{U} \in \mathbb{R}^{n\times K}$ such that $\boldsymbol{U} = \boldsymbol{\Theta}\boldsymbol{X}$ with $\boldsymbol{\Theta} \in \boldsymbol{M}_{n,K}$, $\boldsymbol{X} \in \mathbb{R}^{K\times K}$, let $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{X}})$ be a $(1+\epsilon)$-approximate solution to the K-means problem for $\widehat{\boldsymbol{U}}$. Define $\delta_k = \min_{l\neq k} \|\boldsymbol{X}_{l*} - \boldsymbol{X}_{k*}\|$. If*

$$(16+8\epsilon)\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_F^2 \leq n_k \delta_k^2 \quad \text{for all } k = 1, \dots, K \ , \tag{4}$$

then, there exists a $K \times K$ permutation matrix $\boldsymbol{J}$ such that $\widehat{\boldsymbol{\Theta}}_{G_*} = \boldsymbol{\Theta}_{G_*}\boldsymbol{J}$ where $G_* = \cup_k G_k \setminus S_k$ and S_k satisfies

$$\sum_{k=1}^{K} |S_k|\delta_k^2 \leq 8(2+\epsilon)\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_F^2 \ .$$

Remark. When the norm $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_F$ is small enough in front of the distance between the rows of $\boldsymbol{X}$, then Lemma 3 explicitly bounds the number of ill-classified nodes.

Proof of Lemma 3. Denote $\overline{\boldsymbol{U}} = \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{X}}$. By triangular inequality, we have

$$\|\overline{\boldsymbol{U}} - \boldsymbol{U}\|_F^2 \leq 2\|\overline{\boldsymbol{U}} - \widehat{\boldsymbol{U}}\|_F^2 + 2\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_F^2 \leq (4+2\epsilon)\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_F^2 \ ,$$

where we used that $\overline{\boldsymbol{U}}$ is $(1+\epsilon)$ solution of K-means for $\widehat{\boldsymbol{U}}$. For any $k = 1, \dots, K$, define $S_k := \{i \in G_k : \ \|\overline{\boldsymbol{U}}_{i*} - \boldsymbol{U}_{i*}\| \geq \delta_k/2$, the set of nodes with label k whose corresponding row $\overline{\boldsymbol{U}}_{i*}$ is far from the population value $\boldsymbol{U}_{i*}$. By Markov's inequality,

$$\sum_{k=1}^{K} |S_k|\delta_k^2/4 \leq \sum_{k=1}^{K} \|\overline{\boldsymbol{U}}_{G_k} - \boldsymbol{U}_{G_k}\|_F^2 = \|\overline{\boldsymbol{U}} - \boldsymbol{U}\|_F^2 \leq (4+2\epsilon)\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_F^2 \ .$$

It remains to prove that nodes are correctly classified outside S_k, $k = 1, \dots, K$. By Condition 4, $|S_k|$ is smaller than n_k for all $k = 1, \dots, K$ and the sets $T_k := G_k \setminus S_k$ are therefore non empty. Consider any node i and j with $i \in T_k$ and $j \in T_l$ for some $k \neq l$. We have $\overline{\boldsymbol{U}}_{i*} \neq \overline{\boldsymbol{U}}_{i*}$, otherwise $\max(\delta_k, \delta_l) \leq \|\boldsymbol{U}_{i*} - \boldsymbol{U}_{j*}\| \leq \|\boldsymbol{U}_{i*} - \boldsymbol{U}_{j*}\| < \|\boldsymbol{U}_{i*} - \overline{\boldsymbol{U}}_{i*}\| + \|\boldsymbol{U}_{j*} - \overline{\boldsymbol{U}}_{j*}\| < \delta_k/2 + \delta_l/2$ since $i \in T_k$ and $j \in T_l$. Consequently, $\overline{\boldsymbol{U}}$ has exactly K distinct rows. If i and j belong to same group T_k, then $\overline{\boldsymbol{U}}_{i*} = \overline{\boldsymbol{U}}_{i*}$, otherwide $\overline{\boldsymbol{U}}$ would have more than K distinct rows. In conclusion, $\overline{\boldsymbol{U}}$ correctly classifier (up to permutation) then nodes in T_k, $k = 1, \dots, K$.

□

In order to control the Frobenius norm $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_F$, we rely on the celebrated $\sin(\theta)$ Davis-Kahan inequalities.

Lemma 4 (Davis-Kahan inequality [14]). *Assume that $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is a rank K symmetric matrix with smallest (in absolute value) nonzero singular value γ_n. Let $\boldsymbol{A}$ be any symmetric matrix and $\widehat{\boldsymbol{U}}$, $\boldsymbol{U} \in \mathbb{R}^{n\times K}$ be the K leading eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{P}$, respectively. Then there exists a $K \times K$ orthogonal matrix $\boldsymbol{Q}$ such that*

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F^2 \leq \frac{8K}{\gamma_n^2}\|\boldsymbol{A} - \boldsymbol{P}\|^2$$

Remark: Usual versions of Davis-Kahan inequality express either as $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F^2 \leq \frac{8}{\gamma_n^2}\|\boldsymbol{A} - \boldsymbol{P}\|_F^2$ or $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|^2 \leq \frac{8}{\gamma_n^2}\|\boldsymbol{A} - \boldsymbol{P}\|^2$. In Lemma 4, we combine the latter inequality with the bound $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F^2 \leq K\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|^2$. This is preferred to the first Davis-Kahan inequality because in most relevant setting the Frobenius norm $\|\boldsymbol{A}-\boldsymbol{P}\|_F^2$ is much larger than $K\|\boldsymbol{A}-\boldsymbol{P}\|^2$. Davis-Kahan inequalities are proved using matrix perturbation arguments.

As a consequence of the two above lemmas, we can bound the reconstruction error of the spectral clustering algorithm 1 in terms of the spectral norm $\|\boldsymbol{A} - \boldsymbol{P}\|$, $\|\mathcal{L}(\boldsymbol{A}) - \mathcal{L}(\boldsymbol{P})\|$, or more generally $\|\Psi(\boldsymbol{A}) - \boldsymbol{P}\|$ or $\|\Psi(\boldsymbol{A}) - \mathcal{L}(\boldsymbol{P})\|$.

2.4 Reconstruction bounds

Denote $p_{\max} := \max_{i,j} \boldsymbol{B}_{i,j}$ and $p_{\min} := \min_{i,j} \boldsymbol{B}_{i,j}$. Assume that the procedure *DistanceClustering* computes a 10-approximation of K-means problem using for instance the algorithm of [19][4].

From the above analysis, we need to control the spectral norm of $\boldsymbol{A}-\boldsymbol{P}$ or $\mathcal{L}(\boldsymbol{A})-\mathcal{L}(\boldsymbol{P})$. In this subsection, we collect several concentration inequalities and deduce reconstruction bounds for the corresponding spectral algorithm. The concentration inequalities are stated for the more general inhomogeneous Erdős-Renyi model.

Definition 2 (Inhomogeneous Erdős-Renyi random graph)**.** *Let $\boldsymbol{P}$ be a $n \times n$ symmetric matrix whose entries belong to $[0, 1]$. A random matrix $\boldsymbol{A}$ is distributed according to an Inhomogeneous Erdős-Renyi random graph with parameter $\boldsymbol{P}$ if*

- *$\boldsymbol{A}$ is symmetric*
- *all the above-diagonal entries of $\boldsymbol{A}$ follow independent Bernoulli distributions with parameter $\boldsymbol{P}_{i,j}$.*

2.4.1 Adjacency spectral clustering

The adjacency spectral clustering method corresponds to $\Psi = \mathrm{Id}$ in Algorithm 1.

Proposition 1 ([15, 30, 34])**.** *There exist two positive constants C_0 and C such that the following holds. Let $\boldsymbol{A}$ be the adjacency matrix of an inhomogeneous Erdős-Renyi random graph with parameter $\boldsymbol{P}$. Let $\sigma^2 := \max_{i,j} \boldsymbol{P}_{ij}(1 - \boldsymbol{P}_{ij})$. If $\sigma^2 \geq C_0 \log n/n$, then*

$$P(\|\boldsymbol{A} - \boldsymbol{P}\| \geq C\sigma n^{1/2}) \leq n^{-3} \ .$$

Remark. It is possible to replace n^{-3} by any n^{-k} by changing the constant C and C_0 in the above inequality. Also, this deviation inequality is (up to constants) tight when all the probabilities $\boldsymbol{P}_{ij}$ are of the same order. Classical matrix concentration inequalities such as noncommutative Bernstein inequalities [32] do not allow to recover Proposition 1 (a $\log(.)$ term is missing).

Equipped with this concentration inequality, we arrive at the first reconstruction bound for spectral adjacency clustering.

Theorem 1 ([22])**.** *There exist three positive constants C_0, C_1 and C_2 such that the following holds. Assuming that $p_{\max} \geq C_0 \log(n)/n$ and that*

$$u_n := \frac{Knp_{\max}}{n_{\min}^2 \lambda_{\min}^2(\boldsymbol{B})} \tag{5}$$

[4]The constant 10 is arbitrary. We only need to consider a $(9 + \epsilon)$-approximation.

is smaller than C_1, the adjacency spectral reconstruction $\widehat{Z}$ satisfies $l_*(\widehat{Z}, Z) \leq C_2 u_n$ with probability larger than $1 - n^{-3}$.

Proof. In this proof, C is a numerical constant that may vary from line to line. Denote $\mathrm{Diag}(\boldsymbol{P})$ the diagonal matrix whose diagonal elements coincide with those of $\boldsymbol{P}$. Since $\boldsymbol{A}$ follows an inhomogeneous ErdHos-Renyi with parameter $\boldsymbol{P} - \mathrm{Diag}(\boldsymbol{P})$, Proposition 1 gives

$$\|\boldsymbol{A} - \boldsymbol{P}\| \leq \|\mathrm{Diag}(\boldsymbol{P})\| + \|\boldsymbol{A} - \mathrm{Diag}(\boldsymbol{P})\| \leq C\sqrt{np_{\max}} \ ,$$

with probability larger than $1 - n^{-3}$. By Davis-Kahan inequality (Lemma 4), we arrive at

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{\Theta}\boldsymbol{X}\boldsymbol{Q}\|_F^2 \leq C\frac{Knp_{\max}}{\lambda_K^2(\boldsymbol{P})} \ ,$$

where $\boldsymbol{Q}$ is some orthogonal matrix and $\lambda_K(\boldsymbol{P})$ is the K-th largest (in absolute value) eigenvalue of $\boldsymbol{P}$.

Lemma 5.

$$|\lambda_K(\boldsymbol{P})| \geq n_{\min}|\lambda_{\min}(\boldsymbol{B})| \ .$$

From the above lemma (proved below), it then follows

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{\Theta}\boldsymbol{X}\boldsymbol{Q}\|_F^2 \leq C\frac{Knp_{\max}}{n_{\min}^2\lambda_{\min}^2(\boldsymbol{B})} \ , \tag{6}$$

In order to apply Lemma 3 to $\widehat{\boldsymbol{U}}$, we need to bound the l_2 distance between distinct rows of $\boldsymbol{\Theta}\boldsymbol{X}\boldsymbol{Q}$. By Lemma 1,

$$\delta_k^2 = \frac{1}{n_k} + \inf_{l \neq k}\frac{1}{n_l} \geq \frac{1}{n_k}$$

so that $n_k\delta_k^2 \geq 1$. Gathering (6) with Condition (5), we are in position to apply the reconstruction bound for approximate solutions of K-means (Lemma 3). Up to a permutation, all nodes outside S_k, $k = 1, \dots K$ are therefore correctly classified and S_k satisfies

$$\sum_{k=1}^{K}\frac{|S_k|}{n_k} \leq C\frac{Knp_{\max}}{n_{\min}\lambda_{\min}^2(\boldsymbol{B})}$$

□

Proof of Lemma 5. Up to a permutation of rows and columns of $\boldsymbol{P}$, we may assume that is a block constant matrix with $K \times K$ blocks. Denoting v an eigenvector of $\boldsymbol{P}$ associated to $\lambda_K(\boldsymbol{P})$, the vector v is block constant and writes as $v = (u_1, \dots, u_1, u_2, \dots, u_2, \dots, u_K)^*$ for some $u \in \mathbb{R}^K$ and u_i is repeated n_i times. The vector $\boldsymbol{P}v$ also decomposes as $\boldsymbol{P}v = (w_1, \dots, w_1, \dots, w_K)^*$ for some $w \in \mathbb{R}^K$. By definition of $\boldsymbol{P}$, we have

$$w = \boldsymbol{B}(u \odot \underline{n}) \ ,$$

where $\underline{n} = (n_1, \dots n_K)$ and $\odot$ denotes the coordinate-wise product. Hence,

$$\|\boldsymbol{P}v\|^2 \geq n_{\min}\|w\|^2 \geq n_{\min}\lambda_{\min}^2(\boldsymbol{B})\|u \odot \underline{n}\| \geq n_{\min}^2\lambda_{\min}^2(\boldsymbol{B})\|v\|^2 = n_{\min}^2\lambda_{\min}^2(\boldsymbol{B}) \ ,$$

where we used $\|v\|^2 = \sum_{i=1}^K n_i u_i^2$. This concludes the proof. □

Remark: For sparser graphs ($p_{\max} = O(1/n)$), the adjacency matrix $\boldsymbol{A}$ does not concentrate around $\boldsymbol{P}$ so that the adjacency spectral clustering method performs poorly. The analysis of sparse SBM requires dedicated procedures that will be discussed later. When all the groups are of comparable size ($n_{\min} \approx \frac{n}{K}$), then Condition (5) simplifies as

$$\frac{K^3 p_{\max}}{n\lambda^2_{\min}(\boldsymbol{B})} \leq C_1' \ . \tag{7}$$

The optimality of this bound is discussed in Section 4.

2.4.2 Laplacian spectral clustering

The Laplacian spectral clustering method corresponds to $\Psi = \mathcal{L}$ in Algorithm 1.

Proposition 2 ([29]). *Let $\boldsymbol{A}$ be the adjacency matrix of an Inhomogeneous Erdős-Renyi random graph with $n \times n$ parameter $\boldsymbol{P}$. For any constant $c > 0$ there exists another constant $C = C(c) > 0$ such that the following holds. Let $d := \min_{i=1,\dots,n} \sum_{j=1}^n \boldsymbol{P}_{ij}$. If $d \geq C\log(n)$, then for all $n^{-c} \leq \delta \leq 1/2$,*

$$\mathbb{P}\left[\|\mathcal{L}(\boldsymbol{A}) - \mathcal{L}(\boldsymbol{P})\| \geq 14\sqrt{\frac{\log(4n/\delta)}{d}}\right] \leq \delta \ .$$

Theorem 2. *There exist three positive constants C_0, C_1 and C_2 such that the following holds. Assuming that $p_{\min} \geq C_0 \log(n)/n$ and that*

$$v_n := \frac{Knp^2_{\max}\log(n)}{p_{\min} n^2_{\min}\lambda^2_{\min}(\boldsymbol{B})} \tag{8}$$

is smaller than C_1, the Laplacian spectral reconstruction $\widehat{Z}$ satisfies $l_*(\widehat{Z}, Z) \leq C_2 v_n$ with probability larger than $1 - n^{-3}$.

The proof is similar to that of 1 except that we apply the concentration bound for the Laplacian matrix. The reconstruction error in Theorem 2 is slightly slower than that of Theorem 1 for Adjacency spectral clustering by a multiplicative factor of order $\frac{p_{\max}}{p_{\min}}\log(n)$. As both theorems only provide upper bounds for reconstruction error, it is not clear whether adjacency spectral clustering really outperforms Laplacian spectral clustering. In practice, Laplacian spectral clustering is often preferred as it is reported to produce more stable results.

2.4.3 Regularized Laplacian spectral clustering

When the graph is sparse ($p_{\max} = O(1/n)$), both the adjacency matrix $\boldsymbol{A}$ and the Laplacian matrix $\mathcal{L}(\boldsymbol{A})$ do not concentrate well enough around their population values $\boldsymbol{P}$ and $\mathcal{L}(\boldsymbol{P})$. As a consequence, classical spectral algorithms such as the ones described above do not achieve weak recovery(see e.g. [21]). Nevertheless, slight modifications of the Laplacian matrix allow to get a sharp estimate of $\mathcal{L}(\boldsymbol{P})$. Up to our knowledge, Laplacian regularization has been first introduced in [4].

Given $\tau > 0$, define the regularized adjacency matrix $\boldsymbol{A}_\tau$ is defined by $\boldsymbol{A} + \tau\boldsymbol{J}$. where $\boldsymbol{J}$ is the $n \times n$ matrix whose entries are all equal to one. Similarly, denote $\boldsymbol{P}_\tau = \boldsymbol{P} + \tau\boldsymbol{J}$. Recently, [21] have proved that, even for small τ, $\mathcal{L}(\boldsymbol{A}_\tau)$ concentrates well around $\mathcal{L}(\boldsymbol{P}_\tau)$.

Proposition 3 ([21]). *Let $\boldsymbol{A}$ be the adjacency matrix of an Inhomogeneous Erdős-Renyi random graph with $n \times n$ parameter $\boldsymbol{P}$. Let numbers $d \geq e$, $d_0 > 0$ and α be such that*

$$\max_{i,j} n\boldsymbol{P}_{i,j} \leq d \ , \qquad \min_{j=1,\dots,n} \sum_{i=1}^{n} \boldsymbol{P}_{i,j} \geq d_0, \quad \frac{d}{d_0} \leq \alpha$$

Then for any $r \geq 1$, with probability at least $1 - n^{-r}$ we have

$$\|\mathcal{L}(\boldsymbol{A}_\tau) - \mathcal{L}(\boldsymbol{P}_\tau)\| \leq Cr\alpha^2 \log^3(d) \left(\frac{1}{\sqrt{d}} + \frac{1}{\sqrt{n\tau}} \right) \ ,$$

where $C > 0$ is a numerical constant.

Furthermore, simple calculations give

$$\|\mathcal{L}(\boldsymbol{P}_\tau) - \mathcal{L}(\boldsymbol{P})\| \leq C\alpha^2 \left[\frac{n\tau}{d} + \sqrt{\frac{n\tau}{d}} \right] \ .$$

Combining the above lemma with the general strategy for Spectral clustering, one obtains a reconstruction bound for regularized Laplacian spectral clustering ($\Psi(\boldsymbol{A}) = \mathcal{L}(\boldsymbol{A}_\tau)$ in algorithm 1).

Corollary 1. *There exist three positive constants C_0, C_1 and C_2 such that the following holds. Fix $\tau = \sqrt{np_{\max}}$. If*

$$\omega_n := \frac{Kn^{3/2} \log(np_{\max}) p_{\max}^{11/2}}{n_{\min}^2 \lambda_{\min}^2(\boldsymbol{B}) p_{\min}^4} \tag{9}$$

is smaller than C_1, the regularized Laplacian spectral reconstruction $\widehat{Z}$ satisfies $l_*(\widehat{Z}, Z) \leq C_2\omega_n$ with probability larger than $1 - n^{-3}$.

To have a taste of Corollary 1, consider the following asymptotic regime $\boldsymbol{B} = \boldsymbol{B}_0\gamma_n$ where $\gamma_n = o(1)$ and $\boldsymbol{B}_0$ is fixed. Besides, assume that K is fixed and $n_1 = \ldots = n_K = n/K$. Then, ω_n in (9) simplifies as

$$\omega_n = \frac{(\max_{i,j}(\boldsymbol{B}_0)_{i,j})^{11/2}}{(\min_{i,j}(\boldsymbol{B}_0)_{i,j})^4} \cdot \frac{K^3}{(n\gamma_n)^{1/2}\lambda_{\min}^2(\boldsymbol{B}_0)} \ .$$

According to Corollary 1, regularized Laplacian spectral reconstruction achieves weak recovery for γ_n as small as $c(\boldsymbol{B}_0, K)/n$ where $c(\boldsymbol{B}_0, K) > 0$ only depends on $\boldsymbol{B}_0$ and K. Furthermore, regularized Laplacian spectral reconstruction achieves strong recovery if $n\gamma_n \to \infty$.

For denser graphs ($\gamma_n \gg \log(n)/n$), the reconstruction rate of Corollary 1 is slower than that of Laplacian spectral clustering methods, but a specific analysis of the regularized Laplacian in this regime together with a proper tuning of τ should allow to bridge the gap between the two methods.

2.4.4 Trimmed Adjacency spectral clustering

One can also modify the adjacency spectral method to handle sparse graphs. Before this, let us get see why the adjacency method is failing. Consider an (homogeneous) Erdős-Renyi random graph with probability $\boldsymbol{P} = a/n$ with $a > 0$. For large n, the degree of a node asymptotically follows a Poisson distribution with parameter a. However, the maximum degree of the graph is order $\log(n)/\log\log(n)$. In fact, the presence of these "atypically high-degree" nodes with larger degrees is the main reason for $\|\boldsymbol{A} - \boldsymbol{P}\|$ to be large. This is why [3] have proposed to remove these high degree nodes of the adjacency matrix. This method is usually called trimming.

Given a matrix $\boldsymbol{C}$, some subsets V_1 and V_2 of indices, we denote $\boldsymbol{C}_{V_1 \times V_2}$ the corresponding submatrix of $\boldsymbol{C}$.

Proposition 4 ([12]). *There exist a positive constant C such that the following holds. Let $\boldsymbol{A}$ be the adjacency matrix of an inhomogeneous Erdős-Renyi random graph with parameter $\boldsymbol{P}$. Let σ be a quantity such that $\sigma^2 \leq \boldsymbol{P}_{ij}$. Let V be the set of nodes whose degree is smaller than $20\sigma^2 n$. Then,*

$$P(\|(\boldsymbol{A}-\boldsymbol{P})_{V\times V}\| \geq C\sigma n^{1/2}) \leq n^{-3} \ .$$

Definition 3 (Trimmed spectral clustering). *Fix $\sigma^2 = p_{\max}$ and define $\boldsymbol{A}_{\mathrm{Tri}} := \boldsymbol{A}_{V\times V}$ with V defined as in proposition 4 above. Denote $\tilde{n}_{\mathrm{Tri}} = |V|$, $\tilde{n}_k = |G_k \cap V|$ for $k = 1, \dots, K$, and $\tilde{n}_{\min} = \min_k \tilde{n}_k$. Compute a partition of V by applying the adjacency spectral clustering algorithm to $\boldsymbol{A}_{\mathrm{Tri}}$. Label arbitrarily the nodes outside V.*

Corollary 2. *There exist two positive constants C_1 and C_2 such that the following holds with probability larger than $1 - n^{-3}$. if*

$$\tilde{u}_n := \frac{K\tilde{n}p_{\max}}{\tilde{n}^2_{\min}\lambda^2_{\min}(\boldsymbol{B})} \tag{10}$$

is smaller than C_1, the trimmed spectral reconstruction $\widehat{Z}$ satisfies $l_*(\widehat{Z}, Z) \leq C_2\tilde{u}_n + \frac{n-\tilde{n}_{\mathrm{Tri}}}{n_{\min}}$.

The proof follows the same general strategy as all the previous results. The quantity $\tilde{u}_n$ is random as it depends on the subset V. Using standard concentration inequalities for binomial distribution, one can prove that $n \approx \tilde{n}$ and $\tilde{n}_{\min} \approx n_{\min}$ with large probability as soon as $n_{\min}$ is large in front of $\log(n)$. Under this assumption, the reconstruction bounds of trimmed spectral clustering are (up to the additional term $(n - \tilde{n}_{\mathrm{Tri}})/n_{\min}$) of the same order as the reconstruction bound for adjacency spectral clustering (Theorem 1), except that the bounds are now valid for $p_{\max}$ arbitrarily small. Note that, in the dense case ($p_{\max} \gg \log(n)/n$), $\boldsymbol{A}_{\mathrm{Trim}} = \boldsymbol{A}$ with probability going to one, so that the trimmed spectral clustering is asymptotically equivalent to adjacency spectral clustering. In comparison to the previous section, the reconstructions bounds are somewhat nicer than that of regularized spectral clustering.

Unfortunately, the above method requires the knowledge of $p_{\max}$ (or at least an upper bound of $p_{\max}$). Further work is needed to design adaptive trimmed procedures, but see [13] for results in this direction.

3 Low-rank clustering

Until now, we assumed that the matrix $\boldsymbol{B}$ is full rank. Furthermore, all the reconstruction bounds were depending on the smallest (in absolute values) eigenvalue of $\boldsymbol{B}$. In this section, we improve on this restriction by simply assuming

H.0: The rows of $\boldsymbol{B}$ are distinct.

This assumption is minimal as the label reconstruction problem is not identifiable when **H.0** is not satisfied.

3.1 Algorithm

The general approach analyzed in this section is described in Algorithm 2. For simplicity, take $\Psi = \mathrm{Id}$. Instead of computing the $n \times K$ matrix of eigenvectors of $\boldsymbol{A}$, Algorithm 2 first computes the K-low rank approximation $\widehat{\boldsymbol{A}}_K$ of $\boldsymbol{A}$ defined by

$$\widehat{\boldsymbol{A}}_K \in \arg \min_{\boldsymbol{B},\ \mathrm{Rank}(\boldsymbol{B})=K} \|\boldsymbol{B} - \boldsymbol{A}\|_F^2 \ .$$

The partition $\widehat{G}_1, \dots, \widehat{G}_K$ is obtained by applying a distance-clustering method to the rows of $\widehat{\boldsymbol{A}}_K$. As previously, we assume in the following propositions that *DistanceClustering* computes a 10-approximation of K-means. Up to our knowledge, the idea of using a low-rank approximation of the adjacency matrix has been first considered by [26][5].

Although the rows $\widehat{\boldsymbol{A}}_{k*}$ belong to $\mathbb{R}^n$, the span of these vectors is of dimension less or equal to K, so that the computational complexity of the clustering step in Algorithm 2 is the same as in Algorithm 1.

Algorithm 2 Low-rank projection algorithm

Require: $\boldsymbol{A}$, K, Ψ, *DistanceClustering*

1- Compute the K-low rank approximation $\widehat{\boldsymbol{A}}_K$ of $\Psi(\boldsymbol{A})$.

2- Run *DistanceClustering*$(\widehat{\boldsymbol{A}}_K, K)$.

Output: Partition $\widehat{G}_1, \dots, \widehat{G}_K$ of the nodes.

3.2 Reconstruction bounds

Following the strategy sketched in Lemma 3, we only need to control $\|\widehat{\boldsymbol{A}}_K - \boldsymbol{P}\|_F^2$ (or $\|\widehat{\boldsymbol{A}}_K - \mathcal{L}(\boldsymbol{P})\|_F^2$) to bound the reconstruction loss of $(\widehat{G}_1, \dots, \widehat{G}_K)$. We use the following classical result (see e.g. [16, Ch.6]).

Lemma 6. *Let $\widehat{\boldsymbol{A}}_K$ be a K-low rank approximation of some matrix $\boldsymbol{A}$ and let $\boldsymbol{M}$ be a matrix of rank less or equal to K. Then,*

$$\|\widehat{\boldsymbol{A}}_K - \boldsymbol{M}\|_F^2 \leq 8K\|\boldsymbol{A} - \boldsymbol{M}\|^2 \ . \tag{11}$$

Gathering Lemmas 6 and 3 with the deviation inequalities described in Section 2.4, we can easily adapt the various bounds obtained for spectral clustering to low-rank clustering. Denote $\Delta_{\min}(\boldsymbol{B})$ the minimal l_2 distance between two rows of $\boldsymbol{B}$.

Proposition 5 (Adjacency low-rank reconstruction ($\Psi = \mathrm{Id}$))**.** *There exist three positive constants C_0, C_1 and C_2 such that the following holds. Assuming that $p_{\max} \geq C_0 \log(n)/n$ and that*

$$u_n := \frac{Knp_{\max}}{n_{\min}^2 \Delta_{\min}^2(\boldsymbol{B})} \tag{12}$$

is smaller than C_1, the adjacency low-rank reconstruction $\widehat{Z}$ satisfies $l_*(\widehat{Z}, Z) \leq C_2 u_n$ with probability larger than $1 - n^{-3}$.

In comparison to the spectral adjacency algorithm (Theorem 1), the factor $\lambda_{\min}(\boldsymbol{B})$ is replaced $\Delta_{\min}(\boldsymbol{B})$. In general, $\Delta_{\min}(\boldsymbol{B}) \geq \sqrt{2}\lambda_{\min}(\boldsymbol{B})$. For some matrices $\boldsymbol{B}$, suc as

$$\boldsymbol{B} := \begin{pmatrix} a & b & \frac{a+b}{2} \\ b & a & \frac{a+b}{2} \\ \frac{a+b}{2} & \frac{a+b}{2} & \frac{a+b}{2} \end{pmatrix} \ ,$$

where a and b in $(0, 1)$, there can be a large discrepancy between $\Delta_{\min}(\boldsymbol{B})$ and $\lambda_{\min}(\boldsymbol{B})$.

Exercise: Compute reconstruction bounds for Laplacian ($\Psi = \mathcal{L}$), regularized Laplacian ($\Psi(\boldsymbol{A}) = \mathcal{L}(\boldsymbol{A}_\tau)$ and trimmed adjacency low-rank clustering.

[5] Actually, Mc Sherry [26] uses a more sophisticated method because he is aiming for perfect reconstruction rather than strong reconstruction.

4 A Minimax lower bound

Consider any matrix $\boldsymbol{B}$ satisfying **H.0**. For simplicity, assume that all the groups have the same size

$$|G_1| = |G_2| = \ldots = \frac{n}{K}$$

Define

$$\mathcal{Z} := \left\{ Z \in \{1, \ldots, K\}^n : \ |G_k| = \tfrac{n}{K} \text{ for all } k = 1, \ldots, K \right\} \ ,$$

the corresponding collection of labeling. Given $Z \in \mathcal{Z}$, denote $\mathbb{P}_Z$ the distribution of the adjacency matrix $\boldsymbol{A}$. Below, we give a simple lower bound for minimax risk of reconstruction $\inf_{\widehat{Z}} \sup_{Z \in \mathcal{Z}} \mathbb{E}_Z \left[l_*(\widehat{Z}; Z) \right] \geq C'$.

Proposition 6. *There exists two positive constants C and C' such that the following holds. If*

$$\frac{K p_{\min}}{n \Delta_{\min}^2(\boldsymbol{B})} \geq C$$

then $\inf_{\widehat{Z}} \sup_{Z \in \mathcal{Z}} \mathbb{E}_Z \left[l_*(\widehat{Z}; Z) \right] \geq C'$.

A proof sketch is given in the appendix. Let us compare the minimax lower bound with Proposition 5. In this setting where $n_{\min} = n/K$, (12) is equivalent to

$$\frac{K^3 p_{\max}}{n \Delta_{\min}^2(\boldsymbol{B})} \leq C \ . \tag{13}$$

This condition therefore exhibits the optimal dependency with respect to n and $\Delta_{\min}(\boldsymbol{B})$, at least when the probabilities of connection are of the same order ($p_{\min} \approx p_{\max}$). However, the dependency of (13) on K does not match with the minimax lower bound.

5 Discussion and open questions

Optimal rates for reconstruction. We have seen that spectral-type methods achieve near optimal conditions for weak recovery when K is seen as fixed. When K is seen as a part of the problem, optimal conditions for weak or strong recovery are still unknown, except for some specific matrices $\boldsymbol{B}$. Using a more sophisticated algorithm than 2, Vu [34] has improved the condition (13) by a factor K (times polylog terms) but his condition still does not match the minimax lower bound. It is in fact conjectured that polynomial methods such as spectral clustering or convex relaxations cannot achieve weak recovery under the minimal (ie minimax) conditions. See [11] for partial results in the K-class affiliation model (all diagonal entries of $\boldsymbol{B}$ are equal to $a \in (0,1)$ and all the off-diagonal entries of $\boldsymbol{B}$ are equal to $b \neq a$).

Sharp recovery boundary for the two-class affiliation model. When $n_1 = n_2 = n/2$ and $a > b$[6], it has been conjectured that weak label recovery is possible if and only if

$$\underline{\lim} \quad \frac{n(a-b)^2}{2(a+b)(1-a)} > 1 \ . \tag{14}$$

(a is assumed to be bounded away from one). The impossibility result has been proved [27] in the sparse regime ($a \approx \text{Cste}/n$), but the problem remains open in the dense regime ($a \gg \log(n)/n$). On

[6]The assortative condition $a > b$ is not crucial here. We put it forwar to ease the presentation.

the positive side, the analysis in Section 2 imply that spectral clustering algorithms achieves weak consistency in the dense regime under a condition similar to (14) (with weaker constants). In fact, adjacency spectral clustering achieves weak recovery under the presumably optimal condition (14), but their arguments are more involved. In the sparse regime, regularized Laplacian and trimmed adjacency spectral clustering have been proved (Section 2) to achieve weak recovery under similar conditions to (14) (the constants are again weaker). Tailored procedures (see [9, 20, 24, 28]) achieve the optimal condition (14) but their analysis does not seem to straightforwardly extend to more general SBM models.

Strong recovery versus perfect recovery. In these notes, we reviewed some conditions for strong label recovery, that is the proportion of missclassified nodes is small. A more ambitious goal is to correctly classify all the nodes. In fact, for stochastic block models, it is possible to achieve perfect recovery with a slightly larger signal to noise ratio than needed for strong recovery. This phenomenon has been nicely illustrated in the two-class affiliation model (same setting as in the previous paragraph) where strong recovery is achieved if only if $\frac{n(a-b)^2}{2(a+b)(1-a)} \to \infty$ and perfect recovery is achieved when $\underline{\lim} \frac{n(a-b)^2}{2(a+b)(1-a)\log(n)} > 1$. There is only a logarithmic factor difference between the two conditions. There are two main approaches for achieving perfect recovery: convex relaxations of the maximum likelihood estimator and local improvements of an estimator $\hat{Z}$ achieving strong recovery. The former methods are elegant as they only rely on a single minimization of a semi-definite program [11, 18]. Unfortunately, their definition is, up to our knowledge, restricted to strongly assortative models (all the diagonal elements of $\boldsymbol{B}$ are larger than off-diagonal elements). The latter method are valid in general settings. The general idea is the following: for any $i \in [n]$, assign i to the label $\tilde{Z}_i$ with largest posterior probability given $\boldsymbol{A}$ and $Z_j = \hat{Z}_j$ for all $j \neq i$. In order to ease the analysis, this approach is often combined with sample splitting introduce independence between the first (estimation of $\hat{Z}$) and second (local improvements) step of the procedure. See Abbe et al. [1] for a description in the K-affiliation model. The extension to more general models is explained in [2] and [23].

Extension to degree-corrected and overlapping Stochastic block models. Modifications of spectral clustering algorithm can handle these two extensions [22, 36]. The tensor product method of [6] also achieves strong label recovery of overlapping SBM under suitable conditions. Specified to non-overlapping SBM, reconstruction bounds of [6] are comparable to those obtained in Theorem 1 for spectral clustering.

A Proof sketch for Proposition 6

The proof follows the beaten path of Fano's lemma (see e.g. [31, 35]). By symmetry, we can assume that $\Delta_{\min}(\boldsymbol{B})$ is achieved by the first and second rows of $\boldsymbol{B}$. Define the vector $Z_0 \in \mathcal{Z}$ such that $Z_0 = \{1, \dots, 1, 2, \dots, 2, \dots\}$ and consider the subcollection

$$\mathcal{A} := \big\{ Z \in \mathcal{Z}, \quad Z_i = (Z_0)_i \ \forall i \geq 2\tfrac{n}{K} \text{ and } |G_1 \cap \{1, \dots \tfrac{n}{K}\}| \geq \tfrac{3n}{4K} \big\}$$

of labeling that are close enough to Z_0. For any $Z, Z' \in \mathcal{A}$, the loss $l_*(Z, Z')$ is proportional to the Hamming distance $d_H(.,.)$ between Z and Z'. Then, consider a maximum packing subset $\mathcal{B}$ of $\mathcal{A}$ that is $\epsilon n/K$-separated with respect to the Hamming distance for some small number $\epsilon > 0$. Finally, we bound the Kullback diameter of $\{\mathbb{P}_Z,\ Z \in \mathcal{B}\}$ and lower bound the cardinality of $\mathcal{B}$ by a volumetric argument to apply Fano's lemma.

References

[1] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

[2] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

[3] Noga Alon and Nabil Kahale. A spectral technique for coloring random 3-colorable graphs. *SIAM Journal on Computing*, 26(6):1733–1748, 1997.

[4] Arash A. Amini, Aiyou Chen, Peter J. Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.*, 41(4):2097–2122, 2013.

[5] Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *arXiv preprint arXiv:1406.5647*, 2014.

[6] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

[7] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of euclidean k-means. *arXiv preprint arXiv:1502.03316*, 2015.

[8] Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4):1922–1943, 2013.

[9] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. *arXiv preprint arXiv:1501.06087*, 2015.

[10] Alain Celisse, Jean-Jacques Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.*, 6:1847–1899, 2012.

[11] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

[12] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. *arXiv preprint arXiv:1501.05021*, 2015.

[13] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

[14] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[15] U Feige and E Ofek. Spectral techniques applied to sparse random graphs. *Random Struct. Algorithms*, 27(2):251–275, 2005.

[16] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

[17] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *arXiv preprint arXiv:1411.4686*, 2014.

[18] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

[19] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. In *Proceedings of the eighteenth annual symposium on Computational geometry*, pages 10–18. ACM, 2002.

[20] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zde-

borová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[21] C. M. Le, E. Levina, and R. Vershynin. Sparse random graphs: regularization and concentration of the Laplacian. *ArXiv 1502.03049*, February 2015.

[22] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 2015.

[23] Jing Lei and Lingxue Zhu. A generic sample splitting approach for refined community recovery in stochastic block models. *arXiv preprint arXiv:1411.1469*, 2014.

[24] Laurent Massouli. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, page 694703. ACM, 2014.

[25] C. Matias and S. Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ArXiv 1402.4296*, February 2014.

[26] Frank McSherry. Spectral partitioning of random graphs. In *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)*, pages 529–537. IEEE Computer Soc., Los Alamitos, CA, 2001.

[27] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. Available from http://http://arxiv.org/abs/1202.1499, 2012.

[28] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[29] R. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *ArXiv 0911.0600*, November 2009.

[30] Dan-Cristian Tomozei and Laurent Massoulie. Distributed User Profiling via Spectral Methods. In *Sigmetrics 2010*, volume 38, pages 383–384. ACM Sigmetrics, 2010.

[31] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[32] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[33] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395416, 2007.

[34] Van Vu. A simple svd algorithm for finding hidden partitions. *ArXiv 1404.3918*, April 2014.

[35] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

[36] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting overlapping communities in networks with spectral methods. *arXiv preprint arXiv:1412.3432*, 2014.

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