JMLR: Workshop and Conference Proceedings vol 40:1–33, 2015

Stochastic Block Model and Community Detection in Sparse Graphs: A spectral algorithm with optimal rate of recovery *

Peter Chin SPCHIN@CS.BU.EDU
Boston University

Anup Rao ANUP.RAO@YALE.EDU
Yale University

Van Vu VAN.VU@YALE.EDU
Yale University

Abstract

In this paper, we present and analyze a simple and robust spectral algorithm for the stochastic block model with k blocks, for any k fixed. Our algorithm works with graphs having constant edge density, under an optimal condition on the gap between the density inside a block and the density between the blocks. As a co-product, we settle an open question posed by Abbe et. al. concerning censor block models.

Keywords: Stochastic Block Model, Censor Block Model, Spectral Algorithm

1. Introduction

Community detection is an important problem in statistics, theoretical computer science and image processing. A widely studied theoretical model in this area is the stochastic block model. In the simplest case, there are two blocks V_1, V_2 each of size of n; one considers a random graph generated from the following distribution: an edge between vertices belonging to the same block appears with probability $\frac{a}{n}$ and an edge between vertices across different blocks appear with probability $\frac{b}{n}$, where $a > b > 0$. Given an instance of this graph, we would like to identify the two blocks as correctly as possible. Our paper will deal with the general case of $k \geq 2$ blocks, but for the sake of simplicity, let us first focus on $k = 2$.

For $k = 2$, the problem can be seen as a variant of the well known hidden bipartition problem, which has been studied by many researchers in theoretical computer science, starting with the work of Bui et al. (1987); (see Dyer and Frieze (1989) Boppana (1987) Jerrum and Sorkin (1993) McSherry (2001) and the references therein for further developments). In these earlier papers, a and b are large (at least $\log n$) and the goal is to recover both blocks completely. It is known that one can efficiently obtain a complete recovery if $\frac{(a-b)^2}{a+b} \geq C \frac{\log n}{n}$ and $a, b \geq C \log n$ for some sufficiently large constant C (see, for instance Vu (2014)).

In the stochastic block model problem, the graph is sparse with a and b being constants. Classical results from random graph theory tell us that in this range the graph contains, with high probability, a linear portion of isolated vertices Bollobas (2001). Apparently, there is no way to tell these

* P. Chin is supported by NSF-DMS-1222567 from NSF and FA9550-12-1-0136 from the Air Force Office of Scientific Research. A. Rao is supported by NSF grant CCF-1111257. V. Vu is supported by research grants DMS-0901216 and AFOSAR-FA-9550-09-1-0167 and a grant from Draper Lab.

© 2015 P. Chin, A. Rao & V. Vu.

vertices apart and so a complete recovery is out of question. The goal here is to recover a large portion of each block, namely finding a partition $V_1' \cup V_2'$ of $V = V_1 \cup V_2$ such that V_i and V_i' are close to each other. For quantitative purposes, let us introduce a definition

Definition 1 *A collection of subsets V_1', V_2' of $V_1 \cup V_2$ is γ-correct if $|V_i \cap V_i'| \geq (1-\gamma)n$.*

In Coja-Oghlan (2010), Coja-Oglan proved

Theorem 2 *For any constant $\gamma > 0$ there are constants $d_0, C > 0$ such that if $a, b > d_0$ and $\frac{(a-b)^2}{a+b} > C \log(a+b)$, one can find a γ-correct partition using a polynomial time algorithm.*

Coja-Oglan proved Theorem 2 as part of a more general problem, and his algorithm was rather involved. Furthermore, the result is not yet sharp and it has been conjectured that the $\log$ term is removable[1]. Even when the log term is removed, an important question is to find out the optimal relation between the accuracy γ and the ratio $\frac{(a-b)^2}{a+b}$. This is the main goal of this paper.

Theorem 3 *There are constants C_0 and C_1 such that the following holds. For any constants $a > b > C_0$ and $\gamma > 0$ satisfying*

$$\frac{(a-b)^2}{a+b} \geq C_1 \log \frac{1}{\gamma},$$

we can find a γ-correct partition with probability $1 - o(1)$ using a simple spectral algorithm.

The constants C_0, C_1 can be computed explicitly via a careful, but rather tedious, book keeping. We try not to optimize these constants to simplify the presentation. The proof of Theorem 3 yields the following corollary

Corollary 4 *There are constants C_0 and ϵ such that the following holds. For any constants $a > b > C_0$ and $\epsilon > \gamma > 0$ satisfying*

$$\frac{(a-b)^2}{a+b} \geq 8.1 \log \frac{2}{\gamma},$$

we can find a γ-correct partition with probability $1 - o(1)$ using a simple spectral algorithm.

In parallel to our study, Zhang and Zhou , proving a minimax rate result that suggested that there is a constant $c > 0$

$$\frac{(a-b)^2}{a+b} \leq c \log \frac{1}{\gamma}$$

then one *cannot* recover a γ-correct partition (in expectation), regardless the algorithm.

In order to prove Theorem 3, we design a fast and robust algorithm which obtains a γ-correct partition under the condition $\frac{(a-b)^2}{a+b} \geq C \log \frac{1}{\gamma}$. Our algorithm guarantees γ-correctness with high probability.

We can refine the algorithm to handle the (more difficult) general case of having k blocks, for any fixed number k. Suppose now there are k blocks $V_1, ..., V_k$ with $|V_i| = \frac{n}{k}$ with edge probabilities $\frac{a}{n}$ between vertices within the same block and $\frac{b}{n}$ between vertices in different blocks. As before, a collection of subsets $V_1', V_2', .., V_k'$ of $V_1 \cup V_2 \cup ... \cup V_k$ is γ-correct if $|V_i \cap V_i'| \geq (1-\gamma)\frac{n}{k}$.

1. We would like to thank E. Abbe for communicating this conjecture.

Theorem 5 *There exists constants C_1, C_2, such that if k is any constant as $n \to \infty$ and if*

1. $a > b \geq C_1$
2. $(a-b)^2 \geq C_2 k^2 a \log \frac{1}{\gamma}$,

then we can find a γ-correct partition with probability at least $1 - o(1)$ using a simple spectral algorithm.

We believe that this result is sharp, up to the values of C_1 and C_2; in particular, the requirement $\frac{(a-b)^2}{a} = \Omega(k^2)$ is optimal.

Our method also works (without significant changes) in the case the blocks are not equal, but have comparable sizes (say $cn \geq |V_i| \geq n$ for some constant $c \geq 1$). In this case, the constants C_1, C_2 above will also depend on c. While the emphasis of this paper is on the case a, b are constants, we would like to point out that this assumption is not required in our theorems, so our algorithms work on denser graphs as well.

Let us now discuss some recent works, which we just learned after posting the first version of this paper on arxiv. Mossel informed us about a recent result in Mossel et al. (2013b) which is similar to Theorem 3 (see (Mossel et al., 2013b, Theorem 5.3)). They give a polynomial time algorithm and prove that there exists a constant C such that if $(a-b)^2 > C(a+b)$ and a, b are fixed as $n \to \infty$, then the algorithm recovers an optimal fraction of the vertices. The algorithm in Mossel et al. (2013b) is very different from ours, and uses non-back tracking walks. This algorithm doest not yet handle the case of more than 2 blocks, and its analysis looks very delicate. Next, Guedon sent us Guédon and Vershynin (2014), in which the authors also proved a result similar to Theorem 3, under a stronger assumption $(a-b)^2 \geq C\frac{1}{\gamma^2}(a+b)$ (see Theorem 1.1 and Corollary 1.2 of Guédon and Vershynin (2014)). Their approach relies on an entirely different (semi-definite program) algorithm, which, in turn, was based on Grothendick's inequality. This approach seems to extend to the general $k > 2$ case; however, the formulation of the result in this case, using matrix approximation, is somewhat different from ours (see (Guédon and Vershynin, 2014, Theorem 1.3)). Two more closely related papers have been brought to our attention by the reviewers. In Lelarge et al. (2013), authors have worked out the spectral part of the result in this paper. Also, in a paper posted less than a month prior to this paper, authors of Yun and Proutiere (2014) have also achieved a $\log 1/\gamma$ dependence using similar techniques.

It is remarkable to see so many progresses, using different approaches, on the same problem in such a short span of time. This suggests that the problem is indeed important and rich, and it will be really pedagogical to study the performance of the existing algorithms in practice. We are going to discuss the performance of our algorithm in Sections 3 and 4.

We next present an application of our method to the *Censor Block Model* studied by Abbe et. al. in Abbe et al. (2014). As before, let V be the union of two blocks V_1, V_2, each of size n. Let $G = (V, E)$ be a random graph with edge probability p with incidence matrix B_G and $\boldsymbol{x} = (x_1, ..., x_{2n})$ be the indicator vector of V_2. Let $\boldsymbol{z}$ be a random noise vector whose coordinates z_{e_i} are i.i.d Bernoulli(ϵ) (taking value 1 with probability ϵ and 0 otherwise), where e_i are the edges of G.

Given a noisy observation

$$\boldsymbol{y} = B_G \boldsymbol{x} \oplus \boldsymbol{z}$$

where $\oplus$ is the addition in mod 2, one would like identify the blocks. In Abbe et al. (2014), the authors proved that exact recovery ($\gamma = 0$) is possible if and only if $\frac{np}{\log n} \geq \frac{2}{(1-2\epsilon)^2} + o(\frac{1}{(1-2\epsilon)^2})$ in the limit $\epsilon \to 1/2$. Further, they gave a semidefinite programming based algorithm which succeeds up to twice the threshold. They posed the question of partial recovery ($\gamma > 0$) for sparse graphs. Addressing this question, we show

Theorem 6 *For any given constants $\gamma, 1/2 > \epsilon > 0$, there exists constant C_1, C_2 such that if $np \geq \frac{C_1}{(1-2\epsilon)^2}$ and $p \geq \frac{C_2}{n}$, then we can find a γ-correct partition with probability $1 - o(1)$, using a simple spectral algorithm.*

Let us conclude this section by mentioning a related, interesting, problem, where the purpose is just to do better than a random guess (in our terminology, to find a partition which is $(1/2+\epsilon)$-correct). It was conjectured in Decelle et al. (2011) that this is possible if and only if $(a-b)^2 > (a+b)$. This conjecture has been settled recently by Mossel et. al. Mossel et al. (2012) Mossel et al. (2013a) and Massoulie Massoulié (2013) . Another closely related problem which has been studied in Abbe et al. (2014) Mossel et al. (2014) is about when one can recover at least $1 - o(1)$ fraction of the vertices.

The rest of the paper is organized as follows. In section 2, we describe our algorithm for Theorem 7 and an overview of the proof. The full proof comes in sections 3. In section 4, we show how to modify the algorithm to handle the k block case and prove theorem 5. Finally, in section 5, we prove theorem 6.

2. Two communitites

We first consider the case $k = 2$. Our algorithm will have two steps. First we use a spectral algorithm to recover a partition where the dependence between γ and $\frac{(a-b)^2}{a+b}$ is sub-optimal.

Let A_0 denote the adjacency matrix of a random graph generated from the distribution as in Theorem 7. Let $\bar{A}_0 \stackrel{\text{def}}{=} \mathbb{E}A_0$ and $E_0 \stackrel{\text{def}}{=} A_0 - \bar{A}_0$. Then $\bar{A}_0$ is a rank two matrix with the two non zero eigenvalues $\lambda_1 = a+b$ and $\lambda_2 = a-b$. The eigenvector $\boldsymbol{u}_1$ corresponding to the eigenvalue $a+b$ has coordinates

$$\boldsymbol{u}_1(i) = \frac{1}{\sqrt{2n}}, \text{ for all } i \in V$$

and eigenvector $\boldsymbol{u}_2$ corresponding to the eigenvalue $a-b$ has coordinates

$$\boldsymbol{u}_2(i) = \begin{cases} \frac{1}{\sqrt{2n}} & \text{if } i \in V_1 \\ \frac{-1}{\sqrt{2n}} & \text{if } i \in V_2. \end{cases}$$

Notice that the second eigenvector of $\bar{A}_0$ identifies the partition. We would like to use the second eigenvector of A_0 to approximately identify the partition. Since $A_0 = \bar{A}_0 + E_0$, perturbation theory tells us that we get a good approximation if $\|E_0\|$ is sufficiently small. However, with probability $1 - o(1)$, the norm of E_0 is rather large (even larger than the norm of the main term). In order to handle this problem, we modify E_0 using the auxiliary deletion, at the cost of losing a few large degree vertices.

Spectral Partition.

1. Input the adjacency matrix $A_0, d := a + b$.
2. Zero out all the rows and columns of A_0 corresponding to vertices whose degree is bigger than $20d$, to obtain the matrix A.
3. Find the eigenspace W corresponding to the top two eigenvalues of A.
4. Compute $\boldsymbol{v}_1$, the projection of all-ones vector on to W
5. Let $\boldsymbol{v}_2$ be the unit vector in W perpendicular to $\boldsymbol{v}_1$.
6. Sort the vertices according to their values in $\boldsymbol{v}_2$, and let $V_1' \subset V$ be the top n vertices, and $V_2' \subset V$ be the remaining n vertices
7. Output (V_1', V_2').

Figure 1: Spectral Partition

Let $\bar{A}, A, E$ be the matrices obtained from $\bar{A}_0, A_0, E_0$ after the deletion, respectively. Let $\Delta \stackrel{\text{def}}{=} \bar{A} - \bar{A}_0$; we have

$$\begin{aligned} A &= \bar{A} + E \\ &= \bar{A}_0 + \Delta + E. \end{aligned}$$

The key observation is that $\|E\|$ is significantly smaller than $\|E_0\|$. In the next section we will show that $\|E\| = \mathcal{O}(\sqrt{d})$, with probability $1 - o(1)$, while $\|E_0\|$ is $\Theta(\sqrt{\frac{\log n}{\log\log n}})$, with probability $1-o(1)$. Furthermore, we could show that $\|\Delta\|$ is only $\mathcal{O}(1)$ with probability $1-o(1)$. Therefore, if the second eigenvalue gap for the matrix A_0 is greater than $C\sqrt{d}$, for some large enough constant C, then Davis-Kahan $\sin\Theta$ theorem would allow us to bound the angle between the second eigenvector of $\bar{A}_0$ and A by an arbitrarily small constant. This will, in turn, enable us to recover a large portion of the blocks, proving the following statement

Theorem 7 *There are constants C_0 and C_1 such that the following holds. For any constants $a > b > C_0$ and $\gamma > 0$ satisfying $\frac{(a-b)^2}{a+b} \geq C_1 \frac{1}{\gamma^2}$, then with probability $1 - o(1)$,* **Spectral Partition** *outputs a γ-correct partition.*

Remark 8 *The parameter $d := a + b$ can be estimated very efficiently from the adjacency matrix A. We take this as input for a simpler exposition.*

Step 2 is a further correction that gives us the optimal (logarithmic) dependence between γ and $\frac{(a-b)^2}{a+b}$. The idea here is to use the degree sequence to correct the mislabeled vertices. Consider a mislabeled vertex $u \in V_1' \cap V_2$. As $u \in V_2$, we expect u to have b neighbors in V_1 and a neighbors in V_2. Assume that **Spectral Partition** output V_1', V_2' where $|V_1 \backslash V_i'| \leq 0.1n$, we expect u to have at most $0.9b + 0.1a$ neighbors in V_1' and at least $0.1b + 0.9a$ neighbors in V_2'. As

Partition

1. Input the adjacency matrix $A_0, d := a + b$.
2. Randomly color the edges with Red and Blue with equal probability.
3. Run **Spectral Partition** on Red graph, outputting V_1', V_2'.
4. Run **Correction** on the Blue graph.
5. Output the corrected sets $V_1^{'}, V_2^{'}$.

Figure 2: Partition

$0.1b + 0.9a > \frac{a+b}{2} > 0.9b + 0.1a$, we can correctly reclassify u by thresholding. There are, however, few problems with this argument. First, everything is in expectation. This turns out to be a minor problem; we can use a large deviation result to show that a majority of mislabeled vertices can be detected this way. As a matter of fact, the desired logarithmic dependence is achieved at this step, thanks to the exponential probability bound in the large deviation result.

The more serious problem is the lack of independence. Once **Spectral Partition** has run, the neighbors of u are no longer random. We can avoid this problem using a splitting trick as given in **Partition**. We sample randomly half of the edges of the input graph and used the graph formed by them in **Spectral Partition**. After receiving the first partition, we use the other (random) half of the edges for correction. This doesn't make the two steps completely independent, but we can still prove the stated result.

The sub-routine **Correction** is as follows:

Correction.

1. Input: a partition V_1', V_2' and a Blue graph on $V_1' \cup V_2'$.
2. For any $u \in V_1'$, label u *bad* if the number of neighbors of u in V_2' is at least $\frac{a+b}{4}$ and *good* otherwise.
3. Do the same for any $v \in V_2'$.
4. Correct V_i' be deleting its bad vertices and adding the bad vertices from V_{3-i}'.

Figure 3: Correction

Figure 4 is the density plot of the matrix before and after clustering according to the algorithm described above. We can prove

Lemma 9 *Given a* 0.1*-correct partition* V_1', V_2' *and a Blue graph on* $V_1' \cup V_2'$ *as input to the subroutine* **Correction** *given in figure 3, we get a* γ*-correct partition with* $\gamma = 2\exp(-0.072\frac{(a-b)^2}{a+b})$.

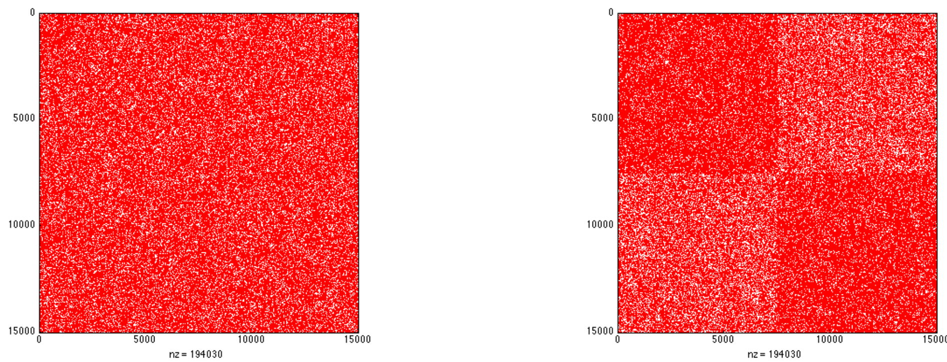

Figure 4: On the left is the density plot of the input (unclustered) matrix with parameters $n = 7500, a = 10, b = 3$ and on the right is the density plot of the permuted matrix after running the algorithm described above. This took less than 3secs in Matlab running on a 2009 MacPro.

3. First step: Proof of Theorem 7

We now turn to the details of the proof. Using the notation in the previous section, we let W be the two dimensional eigenspace corresponding to the top two eigenvalues of A and $\bar{W}$ be the corresponding space of $\bar{A}$. For any two vector subspaces W_1, W_2 of same dimension, we use the usual convention $\sin \angle(W_1, W_2) := \|P_{W_1} - P_{W_2}\|$, where P_{W_i} is the orthogonal projection onto W_i. The proof has two main steps:

1. *Bounding the angle*: We show that $\sin \angle(W, \bar{W})$ is small, under the conditions of the theorem.

2. *Recovering the partition*: If $\sin \angle(W, \bar{W})$ is small, we find an approximate partition which can then improved to find an optimal one.

3.1. Bounding the angle

For the first part, recall that $A = \bar{A}_0 + \Delta + E$. We first prove that $\|\Delta\|$ and $\|E\|$ are small with probability $1 - o(1)$. Bounding $\|\Delta\|$ is easy as it will be sufficient to bound the number of vertices of high degrees. We need the following

Lemma 10 *There exist a constant d_0 such that if $d := a + b \geq d_0$, then with probability $1 - \exp\left(-\Omega(a^{-2}n)\right)$ not more than $a^{-3}n$ vertices have degree $\geq 20d$.*

Note that the proof of the above lemma and other missing proofs in this subsection appear in appendix A.1. If there are at most $a^{-3}n$ vertices with degree $\geq 20d$, then by definition, Δ has at most $2a^{-3}n^2$ non-zero entries, and the magnitude of each entry is bounded by $\frac{a}{n}$. Therefore, its Hilbert-Schmidt norm is bounded by $\|\Delta\|_{HS} \leq \sqrt{2}a^{-1/2}$.

Corollary 11 *For d_0 sufficiently large, with probability $1 - \exp(-\Omega(a^{-3}n))$, $\|\Delta\| \leq 1$.*

Now we address the harder task of bounding $\|E\|$. Here is the key lemma

Lemma 12 *Suppose M is random symmetric matrix with zero on the diagonal whose entries above the diagonal are independent with the following distribution*

$$M_{ij} = \begin{cases} 1 - p_{ij} & \text{w.p.} \quad p_{ij} \\ -p_{ij} & \text{w.p.} \quad 1 - p_{ij} \end{cases}.$$

Let σ be a quantity such that $p_{ij} \leq \sigma^2$ and M_1 be the matrix obtained from M by zeroing out all the rows and columns having more than $20\sigma^2 n$ positive entries. Then with probability $1 - o(1)$, $\|M_1\| \leq C\sigma\sqrt{n}$ for some constant $C > 0$.

Lemma 12 implies

Corollary 13 *There exist constants C_0, C such that if $a > b \geq C_0$, and E is obtained as described before, then we have,*

$$\|E\| \leq C\sqrt{d}$$

with probability $1 - o(1)$.

Now, let $\bar{\boldsymbol{v}}_1, \bar{\boldsymbol{v}}_2$ be eigenvectors of $\bar{A}_0$ corresponding to the largest two eigenvalues $\lambda_1 \geq \lambda_2$ $\boldsymbol{v}_1, \boldsymbol{v}_2$ be eigenvectors of $A = \bar{A}_0 + \Delta + E$ corresponding to the largest two eigenvalues. Further, $\bar{W} :=$ Span$\{\bar{\boldsymbol{v}}_1, \bar{\boldsymbol{v}}_2\}$ and $W :=$ Span$\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$.

Lemma 14 *For any constant $c < 1$, we can choose constants C_2 and C_3 such that such that if $a - b \geq C_2\sqrt{a+b} = C_2\sqrt{d}$ and $a \geq C_3$ then, $\sin(\angle\bar{W}, W) \leq c < 1$ with probability $1 - o(1)$.*

Proof of Lemma 14: Let C_3 be a constant such that if $a \geq C_3$, then theorem 11 holds giving us $\|\Delta\| \leq 1$. From lemma 13 we have that $\|E\| \leq C\sqrt{d}$. The lemma then follows from the Davis-Kahan Davis (1963) Bhatia (1997) bound for matrices $\bar{A}_0$ and A, which gives $\sin(\angle W, \bar{W}) \leq \frac{\|E+\Delta\|}{\lambda_2}$. Therefore, the lemma follows by choosing C_1 big enough.

3.2. Recovery

Given a subspace W satisfying $\sin(\angle\bar{W}, W) \leq c < 1/16$, we can recover a big portion of the vertices. We prove (in appendix A.2) that

Lemma 15 *Given a subspace W satisfying $\sin(\angle\bar{W}, W) \leq c < 1/16$, we can recover a $8c/3$-correct partition.*

Once we have an approximate partition, we can use the Blue edges to boost it in the **Correction** step. We prove (in appendix A.3)

Lemma 16 *Given a 0.1 correct partition V_1', V_2' as input to the* **Correction** *routine in figure 3, the algorithm outputs a γ correction partition with $\gamma = 2\exp(-0.072\frac{(a-b)^2}{a+b})$.*

4. Multiple communities

4.1. Overview

Let us start with the algorithm, which (compared to the algorithm for the case of 2 blocks) has an additional step of random splitting. This additional step is needed in order to recover the partitions. We will start by computing an approximation of the space spanned by the first k eigenvectors of the hidden matrix. However, when $k > 2$, it is not obvious how to approximate the eigenvectors themselves. To handle this problem, we need a new argument that requires this extra step.

Partition

1. Input the adjacency matrix A_0, a, b.
2. Randomly color the edges with Red and Blue with equal probability.
3. Randomly partition V into two subsets Y and Z. Let B be the adjacency matrix of the bipartite graph between Y and Z consisting only of the Red edges, with rows indexed by Z and the columns indexed by Y.
4. Run **Spectral Partition** (figure 6) on matrix B, and get $U'_1, U'_2, ..., U'_k$ as output. This part uses only the Red edges that go between vertices in Y and Z and outputs an approximation to the clustering in $Z = U_1 \cup ... \cup U_k$. Here, $U_i := V_i \cap Z$.
5. Run **Correction** (figure 7) on the Red graph. This procedure only uses the Red edges that are internal to Z and improves the clustering in Z.
6. Run **Merging** (figure 8) on the Blue graph. This part uses only the Blue edges that go between vertices in Y and Z and assigns the vertices in Y to appropriate cluster.

Figure 5: Partition

Since we use different set of edges for each step, we have independence across the steps.

4.2. Details

Step 1 is a spectral algorithm on a portion of the adjacency matrix A_0 as given in figure 6. This will enable us to recover a large portion of the blocks $Z \cap V_1, ..., Z \cap V_k$. We will prove the following statement (appendix B.1)

Theorem 17 *There exists constants C_1, C_2 such that for any fixed integer k the following holds.*

1. $a > b \geq C_1$
2. $\frac{(a-b)^2}{a} \geq C_2 k^2 \frac{1}{\gamma}$, *and*

then we can find a γ-correct partition $U'_1, ..., U'_k$ of Z with high probability using a simple spectral algorithm.

Spectral Partition.

1. Input B (a matrix of dimension $|Z| \times |Y|$), a, b and k.
2. Let Y_1 be a random subset of Y by selecting each element with probability $\frac{1}{2}$ independently and let A_1, A_2 be the sub matrix of B formed by the columns indexed by $Y_1, Y_2 := Y \backslash Y_1$, respectively.
3. Let $d := a + (k-1)b$. Zero out all the rows and columns of A_1 corresponding to vertices whose degree is bigger than $20d$, to obtain the matrix A.
4. Find the space spanned by k left singular vectors of A, say W
5. Let $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ be some $m = 2\log n$ random columns of A_2. For each i, project $\boldsymbol{a}_i - \boldsymbol{a}$ onto W, where $\boldsymbol{a}(j) = \frac{a+b}{2n}$ for all j is a constant vector.
6. For each projected vector, identify the top (in value) $n/2k$ coordinates. Of the $2\log n$ sets so obtained, discard half of the sets with the lowest Blue edge density in them.
7. Of the remaining subsets, identify some k subsets $U'_1, ..., U'_k$ such that $|U'_i \cap U'_j| < 0.2n/2k$, for $i \neq j$.
8. Output $U'_1, ..., U'_k$.

Figure 6: Spectral Partition

Step 2 (figure 7) is a further correction that gives us the optimal (logarithmic) dependence between γ and $\frac{(a-b)^2}{a+b}$. The idea here is to use the degree sequence to correct the mislabeled vertices in Z. Consider a mislabeled vertex $u \in Z \cap V_1$. As $u \in Z \cap V_1$, we expect u to have $a/4$ Red neighbors in $Z \cap V_1$ and $b/4$ Red neighbors in $Z \cap V_i$ for all $i \neq 1$. Assume that **Spectral Partition** output $U'_1, ..., U'_k$ where $|U_1 \backslash U'_1| \leq .1n/2k$, we expect u to have at most $0.9b/4k + 0.1a/4k$ Red neighbors in U'_i and at least $0.1b/4k + 0.9a/4k$ Red neighbors in U'_1. As

$$0.1b/8k + 0.9a/8k > \frac{a+b}{8k} > 0.9b/8k + 0.1a/8k$$

we can correctly reclassify u by thresholding. We can prove (appendix B.2)

Lemma 18 *Given a* 0.1 *correction partition of* $Z = (Z \cap V_1) \cup ... \cup (Z \cap V_k)$ *and the Red graph over* Z*, the sub-routine* **Correction** *given in figure 7 computes a* γ *correct partition with* $\gamma = 2k\exp(-0.04\frac{(a-b)^2}{k(a+b)})$.

Step 3 is to use the clustering information of vertices in Z to label the vertices in Y, and is similar to step 2. We prove (appendix B.3)

Lemma 19 *Given a* 0.1 *correction partition of* $Z = (Z \cap V_1) \cup ... \cup (Z \cap V_k)$ *and the Blue graph over* $Y \cup Z$*, the sub-routine* **Merge** *is given in (figure 8) computes a* γ *correct partition with* $\gamma = 2k\exp(-0.0324\frac{(a-b)^2}{k(a+b)})$.

Combining lemmas 18, 19, we get the stated result.

Correction.

1. Input: A collection of subsets $U_1', ..., U_k' \subset Z$ and a graph on Z.
2. For every $u \in Z$, if $i \in \{1, 2, ..., k\}$ is such that u has maximum neighbors in U_i', then add u to U_i''. Break ties arbitrarily.
3. Output $U_1'', ..., U_k''$.

Figure 7: Correction

Merging.

1. Input: A partition $U_1', ..., U_k'$ of $(Z \cap V_1) \cup (Z \cap V_2) \cup ... \cup (Z \cap V_k)$ and a graph between vertices Y and Z.
2. For all $u \in Y$, label u with **'i'** if the number of neighbors of u in U_i' is at least $\frac{a+b}{8}$. Label the conflicts arbitrarily.
3. Output the label classes as the clusters $V_1^{'}, ..., V_k^{'}$.

Figure 8: Merge

5. Censor Block Model

We first introduce some notations so as to write this problem in a way similar to the other problems in this paper. To simplify the analysis, we make the following assumptions. We assume that there are $|V| = 2n$ vertices, with exactly n of them labeled 1, and the rest labeled 0. As in Abbe et al. (2014), we assume that $G \in G_{2n,p}$ is a graph generated from the Erdos-Renyi model with edge probability p. Since any edge (i, j) appears with probability p, and that $\boldsymbol{z}_e \sim \text{Bernoulli}(\epsilon)$, we have

$$y_{i,j} = \begin{cases} x_i \oplus x_j & w.p. \quad p(1-\epsilon) \\ x_i \oplus x_j \oplus 1 & w.p. \quad p\epsilon \\ 0 & w.p. \quad 1-p \end{cases} .$$

For any $i, j \in V$, let us write $w_{ij} := x_i \oplus x_j$, and $W := (w_{ij})_{ij}$ the associated $2n \times 2n$ matrix.

We note that $\bar{y}_{i,j} := \boldsymbol{E}(y_{i,j}) = p\epsilon + p(1-2\epsilon)w_{i,j}$. Therefore, we can write $y_{i,j} = \bar{y}_{i,j} + \zeta_{i,j}$, where $\zeta_{i,j}s$ are mean zero random variables satisfying $\text{Var}(\zeta_{i,j}) \leq p$. First we note that we can recover the two communities from the eigenvectors of the $2n \times 2n$ matrix $\bar{Y} := (\bar{y}_{i,j}) = p\epsilon I + p(1-2\epsilon)W$. $\bar{Y}$ is a rank 2 matrix with eigenvalues pn and $p(1-2\epsilon)n$, with the corresponding eigenvectors $\boldsymbol{v}_1 = (1, 1, ..., 1)$ and $\boldsymbol{v}_2 = (1, ..., 1, -1, ..., -1)$. if we can find $\boldsymbol{v}_2$, we can identify the two blocks. Let $Y = (y_{i,j})$ and $E = (\zeta_{i,j})$ be $2n \times 2n$ matrices. Algorithm 10 (which is essentially same as algorithm 2) which takes as input the adjacency matrix Y and the edge probability p achieves this when $np \geq \frac{C_2}{(1-2\epsilon)^2}$. More detail appears in appendix C.

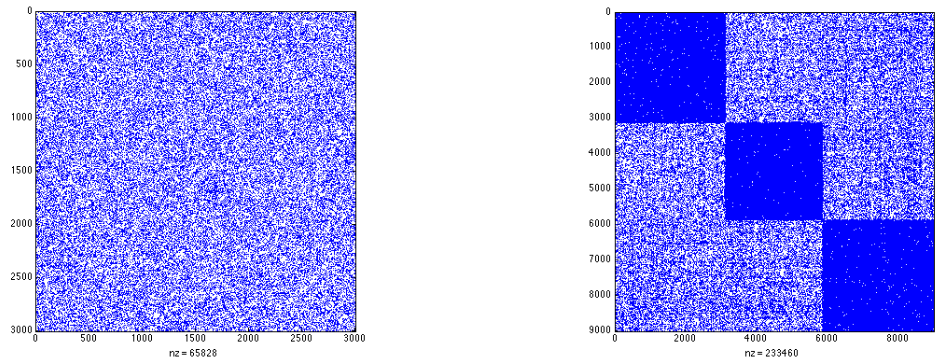

Figure 9: On the left is the density plot of the input (unclustered) matrix with parameters $n = 3000, a = 22, b = 2$ and on the right is the density plot of the permuted matrix after running the algorithm described above. This took less than 1sec in Matlab running on a 2009 MacPro.

Spectral Partition II.

1. Input the adjacency matrix Y, p.
2. Zero out all the rows and columns of Y corresponding to vertices whose degree is bigger than $20pn$, to obtain the matrix Y_0.
3. Find the eigenspace U corresponding to the top two eigenvalues of Y_0.
4. Compute $\boldsymbol{v}_1$, the projection of all-ones vector on to U
5. Let $\boldsymbol{v}_2$ be the unit vector in W perpendicular to $\boldsymbol{v}_1$.
6. Sort the vertices according to their values in $\boldsymbol{v}_2$, and let $V_1' \subset V$ be the top n vertices, and $V_2' \subset V$ be the remaining n vertices
7. Output (V_1', V_2').

Figure 10: Algorithm 3

References

E. Abbe, A. S. Bandeira, and G. Hall. Exact Recovery in the Stochastic Block Model. *ArXiv e-prints*, 1405.3267, May 2014.

Emmanuel Abbe, Afonso S. Bandeira, Annina Bracher, and Amit Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *CoRR*, abs/1404.4749, 2014.

N. Alon and J.H. Spencer. *The Probabilistic Method*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2004. ISBN 9780471653981. URL http://books.google.com/books?id=q3lUjheWiMoC.

Noga Alon and Nabil Kahale. A spectral technique for coloring random 3-colorable graphs. In *SIAM Journal on Computing*, pages 346–355, 1994.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997. ISBN 0387948465.

B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

Ravi B. Boppana. Eigenvalues and graph bisection: An average-case analysis. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, SFCS '87, pages 280–285, Washington, DC, USA, 1987. IEEE Computer Society. ISBN 0-8186-0807-2. doi: 10.1109/SFCS.1987.22. URL http://dx.doi.org/10.1109/SFCS.1987.22.

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities. A nonasymptotic theory of independence, 2013. URL halv3-preprod.archives-ouvertes.fr/hal-00794821.

T.N. Bui, S. Chaudhuri, F.T. Leighton, and M. Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7(2):171–191, 1987. ISSN 0209-9683. doi: 10.1007/BF02579448. URL http://dx.doi.org/10.1007/BF02579448.

Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19:227–284, 3 2010. ISSN 1469-2163. doi: 10.1017/S0963548309990514. URL http://journals.cambridge.org/article_S0963548309990514.

Chandler Davis. The rotation of eigenvectors by a perturbation. *Journal of Mathematical Analysis and Applications*, 6(2):159 – 173, 1963. ISSN 0022-247X. doi: http://dx.doi.org/10.1016/0022-247X(63)90001-5. URL http://www.sciencedirect.com/science/article/pii/0022247X63900015.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011. doi: 10.1103/PhysRevE.84.066106. URL http://link.aps.org/doi/10.1103/PhysRevE.84.066106.

M. E. Dyer and A. M. Frieze. The solution of some random np-hard problems in polynomial expected time. *J. Algorithms*, 10(4):451–489, December 1989. ISSN 0196-6774. doi: 10.1016/0196-6774(89)90001-1. URL http://dx.doi.org/10.1016/0196-6774(89)90001-1.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005. ISSN 1098-2418. doi: 10.1002/rsa.20089. URL http://dx.doi.org/10.1002/rsa.20089.

J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue of random regular graphs. In *Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing*, STOC '89, pages 587–598, New York, NY, USA, 1989. ACM. ISBN 0-89791-307-8. doi: 10.1145/73007.73063. URL http://doi.acm.org/10.1145/73007.73063.

O. Guédon and R. Vershynin. Community detection in sparse networks via Grothendieck's inequality. *ArXiv e-prints*, 1411.4686, November 2014.

Mark Jerrum and G.B. Sorkin. Simulated annealing for graph bisection. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 94–103, Nov 1993. doi: 10.1109/SFCS.1993.366878.

M. Lelarge, L. Massoulie, and Jiaming Xu. Reconstruction in the labeled stochastic block model. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5, Sept 2013. doi: 10.1109/ITW.2013.6691264.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. *CoRR*, abs/1311.3085, 2013.

F. McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537, Oct 2001. doi: 10.1109/SFCS.2001.959929.

E. Mossel, J. Neeman, and A. Sly. Stochastic Block Models and Reconstruction. *ArXiv e-prints*, 1202.1499, February 2012.

E. Mossel, J. Neeman, and A. Sly. A Proof Of The Block Model Threshold Conjecture. *ArXiv e-prints*, 1311.4115, November 2013a.

E. Mossel, J. Neeman, and A. Sly. Belief Propagation, Robust Reconstruction, and Optimal Recovery of Block Models. *ArXiv e-prints*, 1309.1380, September 2013b.

E. Mossel, J. Neeman, and A. Sly. Consistency Thresholds for Binary Symmetric Block Models. *ArXiv e-prints*, 1407.1591, July 2014.

JoelA. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. ISSN 1615-3375. doi: 10.1007/s10208-011-9099-z. URL http://dx.doi.org/10.1007/s10208-011-9099-z.

V. Vu. A simple SVD algorithm for finding hidden partitions. *ArXiv e-prints*, 1404.3918, April 2014.

Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *CoRR*, abs/1412.7335, 2014. URL http://arxiv.org/abs/1412.7335.

Y. Zhang and H. Zhou. Minimax rates of community detection in stochastic block model.

Appendix A. Two communities

A.1. Bounding $\|\Delta\|$ *and* $\|E\|$

Proof of Lemma 10: One can prove Lemma 10 using a standard argument from random graph theory. Consider a set of vertices $X \subset V$ of size $|X| = cn$, where $c < 1$ is a constant. We first bound the probability that all the vertices in this set have degree greater than $20d$.

Let us denote the set of edges on X by $E(X)$ and the set of edges with exactly one end point in X by $E(X, X^c)$. If each degree in X is at least $20d$, then a quick consideration reveals that either $|E(X)| \geq 2cnd$ or $|E(X, X^c)| \geq 8cnd$. The expected number of edges $\mu_{E(X)} := \boldsymbol{E}(|E(X)|)$ satisfies

$$0.25(cn)^2 \frac{a}{n} \leq \mu_{E(X)} \leq 0.5(cn)^2 \frac{a}{n}.$$

Let $\delta_1 := \frac{2}{c} \leq \frac{2cnd}{\mu_{E(X)}}$, then Chernoff bound (see Alon and Spencer (2004) for example) gives

$$\begin{aligned}
\mathbb{P}(|E(X)| \geq cnd) &\leq \left(\frac{\exp(\delta_1 - 1)}{\delta_1^{\delta_1}} \right)^{\mu_{E(X)}} \\
&\leq \exp\left(\left(\frac{2}{c} - 1 - \frac{2}{c} \log\left(\frac{2}{c}\right) \right) 0.25(cn)^2 \frac{a}{n} \right) \\
&\leq \exp\left(-\frac{1}{c} \log\left(\frac{1}{c}\right) 0.25(cn)^2 \frac{a}{n} \right) \quad \text{(for small enough } c) \\
&= \exp\left(-0.25 \log\left(\frac{1}{c}\right) acn \right).
\end{aligned}$$

Similarly, the expected number of edges $\mu_{E(X,X^c)}$ in $E(X, X^c)$ satisfies

$$c(1-c)n^2 \frac{a}{n} \leq \mu_{E(X,X^c)} \leq c(2-c)n^2 \frac{a}{n}.$$

Let $\delta_2 := 4 \leq \frac{8cnd}{\mu_{E(X,X^c)}}$, then by Chernoff bound

$$\begin{aligned}
\mathbb{P}(|E(X, X^c)| \geq 8cnd) &\leq \left(\frac{\exp(\delta_2 - 1)}{\delta_2^{\delta_2}} \right)^{\mu_{E(X,X^c)}} \\
&\leq \exp\left(-c(2-c)an\right).
\end{aligned}$$

Now, if we substitute $c = a^{-3}$ in the above bounds, we get

$$\begin{aligned}
\mathbb{P}(|E(X)| \geq 2cnd) &\leq \exp\left(-0.75 \log(a) a^{-2} n\right) \\
\mathbb{P}(|E(X, X^c)| \geq 8cnd) &\leq \exp\left(-a^{-2} n\right).
\end{aligned}$$

There are at most

$$\binom{2n}{cn} \leq \exp\left(-c\left(\log(\frac{c}{2}) - 1\right) n\right)$$

subsets X of size $|X| = cn$. Substituting $c = a^{-3}$ again, we get

$$\binom{2n}{cn} \leq \exp\left(4a^{-3} \log(a) n\right).$$

The claim follows from the union bound.

Proof of Lemma 12: We start by proving a simpler result.

Lemma 20 *Let M be random symmetric matrix of size n with zero diagonal whose entries above the diagonal are independent with the following distribution*

$$M_{ij} = \begin{cases} 1 - p_{ij} & \text{w.p.} \quad p_{ij} \\ -p_{ij} & \text{w.p.} \quad 1 - p_{ij} \end{cases}.$$

Let $\sigma^2 \geq C_1 \frac{\log n}{n}$ be a quantity such that $p_{ij} \leq \sigma^2$ for all i, j, where C_1 is a constant. Then with probability $1 - o(1)$, $\|M\| \leq C_2 \sigma \sqrt{n}$ for some constant $C_2 > 0$.

Let us address Lemma 20. A weaker bound $C\sigma\sqrt{n \log n}$ follows easily from Alshwede-Winter type matrix concentration results (see Tropp (2012)). To prove the claimed bound, we need to be more careful and follow the ϵ-net approach by Kahn and Szemeredi for random regular graphs in Friedman et al. (1989) (see also Alon and Kahale (1994); Feige and Ofek (2005)).

Consider a $\frac{1}{2}$-net $\mathcal{N}$ of the unit sphere $\mathcal{S}^n$. We can assume $|\mathcal{N}| \leq 5^n$. It suffices to prove that there exists a constant C_2' such that with probability $1 - o(1)$, $|\boldsymbol{x}^T M \boldsymbol{y}| \leq C_2' \sigma \sqrt{n}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{N}$.

For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{N}$, we follow an argument of Kahn and Szemerédi Friedman et al. (1989) and call all pairs (i, j) such that $|x_i y_j| \leq \frac{\sigma}{\sqrt{n}}$ *light* and all remaining pairs *heavy* and denote these two classes by L and H respectively. We have

$$\mathbf{x}^T M \mathbf{y} = \sum_{i,j} x_i M_{ij} y_j = \sum_L x_i M_{i,j} y_j + \sum_H x_i M_{i,j} y_j.$$

We now show that with probability $1 - o(1)$, the last two summands are small in absolute value.

First, let us consider the contribution of light couples. We rewrite $X := \sum_L x_i M_{i,j} y_j$ as $\sum_{(i,j) \in L, i > j} M_{i,j} a_{i,j}$, where

$$a_{i,j} = \begin{cases} x_i y_j + x_j y_i & \text{if} \quad (i,j), (j,i) \in L \\ x_i y_j & \text{if} \quad (i,j) \in L \\ x_j y_i & \text{if} \quad (j,i) \in L \end{cases}$$

By the definition light pairs, $|a_{i,j}| \leq 2\frac{\sigma}{\sqrt{n}}$. Also, since $\mathbf{x}$ and $\mathbf{y}$ are unit vectors, $\sum_{i,j} a_{i,j}^2 \leq 4$. Therefore, by Bernstein's bound (see page 36 in Boucheron et al. (2013) for e.g.)

$$\mathbb{P}(X > t) \leq \exp\left(\frac{-\frac{1}{2}t^2}{4\sigma^2 + \frac{1}{3} 2 \frac{\sigma}{\sqrt{n}} t} \right).$$

Set $t = 10\sigma\sqrt{n}$ and use the union bound (combining with the fact that the net has at most 5^n vectors, we can conclude that with probability at least $1 - \exp(-3n)$, $|\sum_L x_i M_{i,j} y_j| \leq 10\sigma$.

Next we handle the heavy pairs in H. Since $1 \geq \sum_H x_i^2 y_j^2$, the definition of *heavy* implies that $\sum_H |x_i y_j| \leq \frac{\sqrt{n}}{\sigma}$.

Let $A_{i,j} := M_{i,j} + p_{i,j}$, then

$$\sum_H x_i M_{i,j} y_j = \sum_H x_i A_{i,j} y_j - \sum_H p_{ij} x_i y_j.$$

Note that A defines a graph, say G_A, such that A is its adjacency matrix. As $p_{ij} \leq \sigma^2$, we have $\sum_H p_{ij} |x_i y_j| \leq \sigma^2 \frac{\sqrt{n}}{\sigma} = \sigma\sqrt{n}$. We use the following lemma to bound the first term.

Lemma 21 *Let $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ be any graph whose adjacency matrix is denoted by $\widetilde{A}$, and $\boldsymbol{x}, \boldsymbol{y}$ be any two unit vectors. Let $\widetilde{d}$ be such that the maximum degree $\leq c_1\widetilde{d}$. Further, let $\widetilde{d}$ satisfy the property that for any two subsets of vertices $S, T \subset \widetilde{V}$ one of the following holds for some constants c_2 and c_3:*

$$\frac{e(S,T)}{|S||T|\frac{\widetilde{d}}{n}} \leq c_2 \tag{1}$$

$$e(S,T)\log\left(\frac{e(S,T)}{|S||T|\frac{\widetilde{d}}{n}}\right) \leq c_3|T|\log\frac{n}{|T|} \tag{2}$$

then $\sum_H x_i\widetilde{A}_{i,j}y_j \leq \max(16, 8c_1, 32c_2, 32c_3)\sqrt{\widetilde{d}}$. Here $H := \{(i,j) | |x_iy_j| \geq \sqrt{\widetilde{d}}/n\ \}$.

The proof appears in appendix D.

Lemma 22 *Let $\widetilde{d} := \sigma^2 n$. Then with probability $1 - o(1)$, the maximum degree in the graph G_A is $\leq 20\widetilde{d}$ and for any $S, T \subset V$ one of the conditions (1) or (2) holds.*

The two lemmas above guarantee that with probability $1 - o(1)$, $|\sum_H x_iA_{i,j}y_j| \leq C'\sigma\sqrt{n}$ for some constant C'.

Proof The bound on the maximum degree follows from the Chernoff bound. We have that

$$A_{ij} = \begin{cases} 1 & \text{w.p.} \quad p_{ij} \\ 0 & \text{w.p.} \quad 1 - p_{ij} \end{cases}.$$

Consider a particular vertex k and let $X = \sum_i A_{ik}$ be the random variable denoting the number of edges incident on it. We have that

$$\mu = \boldsymbol{E}X = \sum_i p_{ik} \leq \sigma^2 n.$$

For any $l \geq 4$, Chernoff bound (see Alon and Spencer (2004)) implies that

$$\begin{aligned}\mathbb{P}(X > l\sigma^2 n) &\leq \exp\left(-\frac{\sigma^2 nl\ln l}{3}\right) \\ &\leq \exp\left(-\frac{l\log n}{3}\right).\end{aligned}$$

Applying this with $l = 20$, and taking a union bound over all the vertices, we can bound the maximum degree by $20\sigma^2 n$. Now let $S, T \subset V$ be any two subsets. Let $X := e(S,T)$ be the number of edges going between S and T. We have $\boldsymbol{E}X \leq \sigma^2|S||T|$. If $|T| \geq \frac{n}{e}$, then since the maximum degree is $\leq 20\sigma^2 n$, we have $e(S,T) \leq |S|20\sigma^2 n \leq 20e\sigma^2|S||T|$, giving us 1 in this case. Therefore, we can assume $|T| \leq \frac{n}{e}$. By Chernoff bound, it follows that for any $l \geq 4$,

$$\mathbb{P}(e(S,T) > l\sigma^2|S||T|) \leq \exp\left(-\frac{l\ln(l)\sigma^2|S||T|}{3}\right).$$

Let l' be the smallest number such that $l' \ln(l') \geq \frac{21|T|}{\sigma^2|S||T|} \log\left(\frac{n}{|T|}\right)$. As in Feige and Ofek (2005), if we choose $l = \max(l', 4)$, we can bound the above probability by $\exp\left(-\frac{l\ln(l)\sigma^2|S||T|}{3}\right)\binom{n}{|S|}\binom{n}{|T|} \leq \frac{1}{n^3}$. Therefore, by the union bound we get that with probability $1 - o(1)$ for all subsets S, T, and

$$e(S,T) \leq \max(l', 4)\sigma^2|S||T|.$$

This implies that one of the conditions 1 or 2 holds with probability $1 - o(1)$.

■

Proof of Lemma 12: Now we are ready to prove Lemma 12 by modifying the previous proof. We again handle the light couples and the heavy couples separately, but need to make a modification to the argument for the light couples.

Since we zero out some rows and columns of M to obtain M_1, we first bound the norm of the matrix M_0, obtained from M by zeroing out a set S of rows and the corresponding columns. Next, we take a union bound over all choices of S. For a fixed S, lemma 20 implies that with probability at least $1 - \exp(-3n)$, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{N}_{1/2}$, $|\sum_L x_i(M_0)_{ij}y_j| \leq 10\sigma\sqrt{n}$. Since there are at most $2^n = \exp(n\ln 2)$ choices for S, we can apply a union bound to show that with probability at least $1 - \exp(-(3 - \ln 2)n)$, $|\sum_L x_i(M_1)_{ij}y_j| \leq 10\sigma\sqrt{n}$.

The proof for the heavy couples goes through without any modifications. We just have to verify that the conditions of lemma 21 are met. Firstly, the adjacency matrix A_1 obtained from M_1 has bounded degree property by the definition of M_1. Now we note that only for the case of $|S| \leq |T| \geq \frac{n}{e}$ did we need that the maximum degree was bounded. So for any $|S| \leq |T| < \frac{n}{e}$, the discrepancy properties (1) or (2) holds for A_1, since zeroing out rows and columns can only decrease the edge count across sets of vertices. In the case $|T| \geq \frac{n}{e}$, like before we can show that (1) holds for A_1 since the degrees are bounded. ■

Now, to bound the norm of matrix E, we just appeal to 12. Suppose $a > b \geq C_0$, for a large enough constant C_0 to be determined later. Since $A = \bar{A}_0 + \Delta + E$ and we have bounded Δ, it remains to bound $\|E\|$. Note that

$$(E_0)_{ij} = \begin{cases} 1 - \frac{a}{n} & \text{w.p.} \quad \frac{a}{n} \\ -\frac{a}{n} & \text{w.p.} \quad 1 - \frac{a}{n} \end{cases}$$

if i, j belongs to the same community and

$$(E_0)_{ij} = \begin{cases} 1 - \frac{b}{n} & \text{w.p.} \quad \frac{b}{n} \\ -\frac{b}{n} & \text{w.p.} \quad 1 - \frac{b}{n} \end{cases}$$

if i, j belongs to different communities. Since $a > b$, for all i, j we have that

$$\text{Var}((E_0)_{ij}) \leq \frac{a}{n}\left(1 - \frac{a}{n}\right) \leq \frac{d}{n}.$$

A.2. Recovery

Now we focus on the second step in the proof, namely the recovery of the blocks once the angle condition is satisfied.

Lemma 23 *If* $\sin(\angle \bar{W}, W) \leq c \leq \frac{1}{4}$, *then we can find a vector* $\boldsymbol{v} \in W$ *such that* $\sin(\angle \boldsymbol{v}, \bar{\boldsymbol{v}}_2) \leq 2\sqrt{c}$.

Proof Let $P_{\bar{W}}, P_W$ be the orthogonal projection operators on to the subspaces $\bar{W}, W$ respectively. From the angle bound for the subspaces, we have that

$$\|P_{\bar{W}} - P_W\|_2 \leq c.$$

The vector we want is obtained as follows. We first project $\bar{\boldsymbol{v}}_1$ on to W, and then find the unit vector orthogonal to the projection in W. We will now prove that the vector so obtained satisfies the bound stated in the lemma. Since $\bar{\boldsymbol{v}}_1, \bar{\boldsymbol{v}}_2 \in \bar{W}$, we have that $\|P_W \bar{\boldsymbol{v}}_i - \bar{\boldsymbol{v}}_i\|_2 \leq c$ for $i = 1, 2$. Let us define $\boldsymbol{u}_i := P_W \bar{\boldsymbol{v}}_i$ and $\mathbf{x}_i := \boldsymbol{u}_i - \bar{\boldsymbol{v}}_i$ (note that $\|\mathbf{x}_i\| \leq c$) for $i = 1, 2$. We will now show that the vector $\boldsymbol{v} \in W$ perpendicular to $\boldsymbol{u}_1$ is close to $\bar{\boldsymbol{v}}_2$. Let $\boldsymbol{u}_\perp = \boldsymbol{u}_2 - \frac{\boldsymbol{u}_1^T \boldsymbol{u}_2}{\|\boldsymbol{u}_1\|^2} \boldsymbol{u}_1$, it is then clear that $\|\boldsymbol{u}_\perp\| \leq 1$. Note that $|\boldsymbol{u}_1^T \boldsymbol{u}_2| = |\bar{\boldsymbol{v}}_1^T \mathbf{x}_2 + \bar{\boldsymbol{v}}_2^T \mathbf{x}_1 + \mathbf{x}_1^T \mathbf{x}_2| \leq 2c + c^2$. We have,

$$\begin{aligned}
\boldsymbol{u}_\perp^T \bar{\boldsymbol{v}}_2 &= \boldsymbol{u}_2^T \bar{\boldsymbol{v}}_2 - \frac{(\boldsymbol{u}_1^T \boldsymbol{u}_2)(\bar{\boldsymbol{v}}_2^T \boldsymbol{u}_1)}{\|\boldsymbol{u}_1\|^2} \\
\left|\boldsymbol{u}_\perp^T \bar{\boldsymbol{v}}_2\right| &\geq 1 - c - \frac{(2c + c^2)c}{(1-c)^2} \\
&\geq 1 - 2c.
\end{aligned}$$

The last inequality holds when $c \leq \frac{1}{4}$. Therefore, it holds that for a unit vector $\boldsymbol{v} \perp \boldsymbol{u}_1$,

$$|\boldsymbol{v}^T \bar{\boldsymbol{v}}_2| \geq \left|\boldsymbol{u}_\perp^T \bar{\boldsymbol{v}}_2\right| \geq 1 - 2c.$$

This gives $\sin(\angle \boldsymbol{v}, \bar{\boldsymbol{v}}_2) \leq \sqrt{1 - (1-2c)^2} \leq 2\sqrt{c}$. ■

Lemmas 14 and 23 together give

Corollary 24 *For any constant* $c < 1$, *we can choose constants* C_2 *and* C_3 *in lemma 14 and find a vector* $\boldsymbol{v}$ *such that* $\sin(\angle \bar{\boldsymbol{v}}_2, \boldsymbol{v}) \leq c < 1$ *with probability* $1 - o(1)$.

We now can conclude the proof of our theorem using the following deterministic fact.

Lemma 25 *If* $\sin(\angle \bar{\boldsymbol{v}}_2, \boldsymbol{v}) < c \leq 0.5$, *then we can identify at least a* $(1 - \frac{4}{3}c^2)$ *fraction of vertices from each block correctly.*

Proof of Lemma 25: Let us define two sets of vertices, $V_1' = \{i | \boldsymbol{v}(i) > 0\}$ and $V_2' = \{i | \boldsymbol{v}(i) < 0\}$. One of the sets will have less than or equal to $\frac{n}{2}$ vertices, let us assume without loss of generality that $|V_1'| \leq \frac{n}{2}$. Writing $\boldsymbol{v} = c_1 \bar{\boldsymbol{v}}_2 + \mathbf{err}$, for a vector $\mathbf{err}$ perpendicular to $\bar{\boldsymbol{v}}_2$ and $\|\mathbf{err}\| < c$. We also have $c_1 > \sqrt{1 - c^2}$. Since $\|\mathbf{err}\| < c$, not more than $\frac{c^2}{1-c^2} n$ coordinates of $\mathbf{err}$ can be bigger than $\frac{\sqrt{1-c^2}}{\sqrt{n}} < \frac{c_1}{\sqrt{n}}$. Since $\boldsymbol{v} = c_1 \bar{\boldsymbol{v}}_2 + \mathbf{err}$ at least $1 - \frac{c^2}{1-c^2} > 1 - \frac{4}{3}c^2$ (since $c \leq 0.5$) fraction of vertices with $\bar{\boldsymbol{v}}_2(i) = \frac{1}{\sqrt{n}}$ will have $\boldsymbol{v}(i) > 0$. Therefore, we get that there are at least $(1 - \frac{4}{3}c^2)n$ vertices belonging to the first block. ■

A.3. Proof of lemma 9

We will use the following large deviation result (see page 36 in Boucheron et al. (2013) for e.g.) repeatedly

Lemma 26 *(Chernoff) If X is a sum of n iid indicator random variables with mean at most $\rho \leq 1/2$, then for any $t > 0$*

$$\max\{\mathbb{P}(X \geq \boldsymbol{E}X + t), \mathbb{P}(X \leq \boldsymbol{E}X - t)\} \leq \exp\left(-\frac{t^2}{2\,\boldsymbol{Var}X + t}\right) \leq \exp\left(-\frac{t^2}{2n\rho + t}\right).$$

In the Red graph, the edge densities are $a/2n$ and $b/2n$, respectively. By Theorem 7, there is a constant C such that if $\frac{(a-b)^2}{a+b} \geq C$ then by running **Spectral Partition** on the Red graph, we obtain, with probability $1 - o(1)$ two sets V_1' and V_2', where

$$|V_i \backslash V_i'| \leq .1n.$$

In the rest, we condition on this event, and the event that the maximum Red degree of a vertex is at most $\log^2 n$, which occurs with probability $1 - o(1)$.

Now we use the Blue edges. Consider $e = (u, v)$. If e is not a red edge, and $u \in V_i, v \in V_{3-i}$, then e is a Blue edge with probability

$$\mu := \frac{b/2n}{1 - \frac{b}{2n}}.$$

Similarly, if e is not a Red edge, and $u, v \in V_i$, then e is a Blue edge with probability

$$\tau := \frac{a/2n}{1 - \frac{a}{2n}}.$$

Thus, for any $u \in V_i' \cap V_i$, the number of its Blue neighbors in V_{3-i}' is at most

$$S(u) := \sum_{i=1}^{.9n} \xi_i^u + \sum_{j=1}^{.1n} \zeta_j^u$$

where ξ_i^u are iid indicator variables with mean μ and ζ_j^u are iid indicator variables with mean τ.

Similarly, for any $u \in V_1' \cap V_2$, the number of its Blue neighbors in V_2' is at least

$$S'(u) := \sum_{i=1}^{.9n - d(u)} \zeta_i^u + \sum_{j=1}^{.1n} \xi_j^u,$$

where $d(u) = \log^2 n$ is the Red degree of u.

After the correction sub-routine, a vertex u in the (corrected) set V_1' is misclassified if

- $u \in V_1' \cap V_1$ and $S_u \geq \frac{a+b}{4}$.
- $u \in V_1' \cap V_2$ and $S_u' \leq \frac{a+b}{4}$

Let ρ_1, ρ_2 be the probability of the above events. Then the number of misclassified vertices in the (corrected) set V_1' is at most

$$M := \sum_{k=1}^{n} \Gamma_k + \sum_{l=1}^{0.1n} \Lambda_l$$

where Γ_k are iid indicator random variables with mean ρ_1 and Λ_l are iid indicator random variables with mean ρ_2.

The rest is a simple computation. First we use Chernoff bound to estimate ρ_1, ρ_2. Consider

$$\rho_1 := \mathbb{P}\left(S(u) \geq \frac{a+b}{4}\right).$$

By definition, we have

$$\begin{aligned} \boldsymbol{E}S(u) &= 0.9n\mu + 0.1n\tau \\ &= 0.9n(\frac{b/2n}{1-\frac{b}{2n}}) + 0.1n(\frac{a/2n}{1-\frac{a}{2n}}) \\ &= 0.9\frac{b}{2} + 0.1\frac{a}{2} + 0.9\frac{b}{2}(\frac{1}{1-b/2n} - 1) + 0.1\frac{a}{2}(\frac{1}{1-a/2n} - 1). \end{aligned} \tag{3}$$

Set

$$t := \frac{a+b}{4} - \boldsymbol{E}S(u),$$

we have

$$t = 0.2(a-b) - 0.9\frac{b}{2}(\frac{1}{1-b/2n} - 1) - 0.1\frac{a}{2}(\frac{1}{1-a/2n} - 1) \geq 0.2(a-b) - 0.9\frac{b}{2}\frac{b}{n} - 0.1\frac{a}{2}\frac{a}{n} \geq 0.19(a-b),$$

for any sufficiently large n.

Applying Chernoff's bound, we obtain

$$\rho_1 \leq \exp(-\frac{(0.19(a-b))^2}{2(0.9n\mu + .1n\tau) + 0.19(a-b)}).$$

By (4), one can show that $2(0.9n\mu + .1n\tau) + 0.19(a-b) = 0.71b + 0.29a + o(1) \leq \frac{a+b}{2}$. It follows that

$$\rho_1 \leq \exp(-0.072\frac{(a-b)^2}{a+b}).$$

By a similar argument, we obtain the same estimate for ρ_2 (the contribution of the term $d(u) \leq \log^2 n$ is negligible). Thus, we can conclude that

$$\boldsymbol{E}M \leq 1.1n\exp(-0.072\frac{(a-b)^2}{a+b}).$$

Applying Chernoff's with $t := 0.9n\exp(-0.072\frac{(a-b)^2}{a+b})$, we conclude that with probability $1-o(1)$

$$M \leq \boldsymbol{E}M + t = 2n\exp(-0.072\frac{(a-b)^2}{a+b}).$$

This implies that with probability $1-o(1)$,

$$|V_1'\backslash V_1| \leq 2n\exp(-0.072\frac{(a-b)^2}{a+b}).$$

By symmetry, the same conclusion holds for $|V_2'\backslash V_2|$.

Set

$$\gamma := 2\exp(-0.072\frac{(a-b)^2}{a+b}),$$

we have, for $i = 1, 2$

$$|V_i \cap V_i'| = n - |V_i \cap V_{3-i}'| = n - |V_{3-i}'\backslash V_{3-i}| \geq n(1-\gamma).$$

This shows that the output V_1', V_2' form a γ-correct partition, with γ satisfying

$$\frac{(a-b)^2}{a+b} = \frac{1}{0.072}\log\frac{2}{\gamma} \approx 13.89\log\frac{2}{\gamma},$$

proving our claim.

Proof of Corollary 4: Notice that in the analysis of **Spectral Partition**, we only require $\frac{(a-b)^2}{a+b} \geq C$ for a sufficiently large constant C (so γ does not appear in the bound). In the analysis of **Correction**, we require $\frac{(a-b)^2}{a+b} \geq 13.89\log\frac{2}{\gamma}$, as shown above. If $\gamma < \epsilon$ for a sufficiently small ϵ, this assumption implies the first. Thus, Corollary holds with assumption $\frac{(a-b)^2}{a+b} \geq 13.89\log\frac{2}{\gamma}$.

The constant 13.89 comes from the fact that the partition obtained from **Spectral Partition** is .1-correct. If one improves upon .1, one improves 13.89. In particular, there is a constant δ such that if the first partition is δ-correct, then one can improve 13.89 to 8.1 (or any constant larger than 8–which is the limit of the method, for that matter).

Appendix B. Multiple communities

We say the splitting is 'perfect' if we have $|Y_1 \cap V_i| = \frac{n}{4k} = |Y_2 \cap V_i|$ for $i = 1, .., k$. We will assume the splittings are perfect in the proofs for a simpler exposition. Though the splitting will almost always not be perfect, and there will just be a $o(1)$ error term that we have to carry throughout to be precise. The bounds we give will all be still be essentially the same.

B.1. Proof of theorem 17

To analyze this algorithm, we use the machinery developed so far combined with some ideas from Vu (2014). We consider the stochastic block model with k blocks of size n, where k is a fixed constant as n grows. This is a graph $V = V_1 \cup V_2 \cup ... \cup V_k$ where each $|V_i| = n/k$ and for $u \in V_i, v \in V_j$:

$$\mathbb{P}[(u,v) \in E] = \begin{cases} a/n & \text{if } i = j \\ b/n & i \neq j \end{cases}$$

We can write, as before

$$\begin{aligned} A &= \bar{A} + E \\ &= \bar{A}_1 + \Delta + E, \end{aligned}$$

where $\bar{A}$, $\bar{A}_1$ are the expected matrices, and Δ is matrix containing the deleted rows and columns. Let $\bar{W}$ be the span of the k left singular vectors of $\bar{A}_1$ We can bound $\|\Delta\| \leq 1$ by bounding the number of high degree vertices as we did before. E is given by

$$E_{u,v} = \begin{cases} 1 - \frac{a}{n} & \text{w.p.} \quad \frac{a}{n} \\ -\frac{a}{n} & \text{w.p.} \quad 1 - \frac{a}{n} \end{cases}$$

if $u, v \in V_i \cap Y_1$ for some $i \in 1, .., k$ and

$$E_{u,v} = \begin{cases} 1 - \frac{b}{n} & \text{w.p.} \quad \frac{b}{n} \\ -\frac{b}{n} & \text{w.p.} \quad 1 - \frac{b}{n} \end{cases}$$

if $u \in V_i \cap Y_1$ and $v \in V_j \cap Y_1$ for $i \neq j$. Since $\sigma^2 := \frac{a}{n} \geq \text{Var}(E_{u,v})$, corollary 12 applied to $\bar{A}_1 - A$ gives the following result.

Lemma 27 *There exists a constant C such that $\|E\| \leq C\sqrt{a+b}$ with probability $1 - o(1)$.*

It is not hard to show that the rank of the matrix $\bar{A}_1$ is k, and its least non-trivial singular value is $\sigma_k(\bar{A}_1) = \frac{a-b}{k}$. This fact, combined with lemma 27 and an application of Davis-Kahan bound gives

Lemma 28 *For any $c > 0$, there exists constants C_1, C_2 such that if $(a - b) > C_1 k^2 a$ and $a > b \geq C_2$, then $\sin\angle(\bar{W}, W) \leq c$ with probability $1 - o(1)$.*

We pick $m = 2\log n$ indices uniformly randomly from Y_2 and project the corresponding columns from the matrix B. Let $\bar{\boldsymbol{a}}_{i_1}, ..., \bar{\boldsymbol{a}}_{i_m}$ and $\boldsymbol{e}_{i_1}, ..., \boldsymbol{e}_{i_m}$ be the corresponding columns of $\bar{A}_1$ and E, respectively. For a subspace W_0, let P_{W_0} be the projection on to the space W_0. Note that if vertex $i \in V_{n_i} \cap Y_2$, then

$$\bar{\boldsymbol{a}}_i(j) = \begin{cases} \frac{a}{n} & \text{if } j \in V_{n_i} \cap Z \\ \frac{b}{n} & \text{otherwise} \end{cases} .$$

We let the vector $\bar{\boldsymbol{a}}$ be

$$\bar{\boldsymbol{a}}(j) := \begin{cases} \frac{a+b}{2n} & \text{if } j \in Z \end{cases} .$$

and $\boldsymbol{b}_i = \bar{\boldsymbol{a}}_i - \bar{\boldsymbol{a}}$. We therefore have

$$\boldsymbol{b}_i(j) = \begin{cases} \frac{a-b}{2n} & \text{if } j \in V_{n_i} \cap Z \\ \frac{b-a}{2n} & \text{otherwise} \end{cases} .$$

Since both $\bar{\boldsymbol{a}}_i$, $\bar{\boldsymbol{a}}$ are in the column span of $\bar{A}_1$, we have for all i

$$\boldsymbol{b}_i = P_{\bar{W}} \boldsymbol{b}_i.$$

We also note that $\|\boldsymbol{b}_i\| = \frac{(a-b)}{2\sqrt{2n}}$. Therefore, if we can recover $\boldsymbol{b}_i$, we can identify the set $V_{n_i} \cap Z$. We now argue that we can recover $\boldsymbol{b}_i$ approximately. Since $\boldsymbol{a}_i - \bar{\boldsymbol{a}} = \boldsymbol{b}_i + \boldsymbol{e}_i$, we have

$$\begin{aligned} P_W(\boldsymbol{a}_i - \bar{\boldsymbol{a}}) &= P_W \boldsymbol{b}_i + P_W \boldsymbol{e}_i \\ &= P_{\bar{W}} \boldsymbol{b}_i + P_W \boldsymbol{e}_i + \boldsymbol{err}_i \\ &= \boldsymbol{b}_i + P_W \boldsymbol{e}_i + \boldsymbol{err}_i, \end{aligned}$$

where $\boldsymbol{err}_i = (P_W - P_{\bar{W}})\, \boldsymbol{b}_i$. Since $\sin \angle(\bar{W}, W) \leq \delta_1$, we have for any unit vector $\boldsymbol{v}$, $\|P_W \boldsymbol{v} - P_{\bar{W}} \boldsymbol{v}\| \leq \delta_1$, which in turn implies for all i

$$\|\boldsymbol{err}_i\| \leq \delta_1 \|\boldsymbol{b}_i\|.$$

Therefore, it is enough to bound $\|P_W \boldsymbol{e}_i\|$. We recall that k is a constant that does not depend on n. W is k dimensional space giving $\mathbb{E}\|P_W \boldsymbol{e}_i\|^2 \leq k\sigma^2$. By Markov's inequality, it follows that

$$\mathbb{P}(\|P_W \boldsymbol{e}_i\| > 2\sigma k^{1/2}) \leq \frac{1}{4}.$$

By a simple application of Chernoff bound, we have

Lemma 29 *With probability at least $1 - o(1)$, at least $m/2$ of the vectors $\boldsymbol{e}_{i_1}, ..., \boldsymbol{e}_{i_m}$ satisfy*

$$\left\|P_W \boldsymbol{e}_{i_j}\right\| < 2\sigma k^{1/2}.$$

Let $m_1 \geq m/2$ denote the number of such vectors, hence referred to as good vectors. To avoid introducing extra notation, let us say $\boldsymbol{e}_{i_1}, .., \boldsymbol{e}_{i_{m_1}}$ are the good vectors and the corresponding indices as good indices. Note that $\sigma \leq \frac{\sqrt{a}}{\sqrt{n}}$. For any $\delta_2 > 0$, there exists a big enough constant C_1 such that if $(a-b) > C_1\sqrt{ka}$, we have that $2\sigma k^{1/2} \leq \delta_2 \left\|\boldsymbol{b}_{i_j}\right\|$ whenever i_j is good. Therefore

Lemma 30 *Given any $\delta > 0$, there exists constants C_1, C_2 such that the following holds. If $(a - b) > C_1\sqrt{ka}$ and $a \geq b \geq C_2$, then for all good indices i_j, it holds that $\left\|P_W(\boldsymbol{a}_{i_j} - \bar{\boldsymbol{a}}) - \boldsymbol{b}_{i_j}\right\| \leq \delta \left\|\boldsymbol{b}_{i_j}\right\|$.*

Let U'_{i_j} be the top $n/2k$ coordinates of the projected vector $\left\|P_W(\boldsymbol{a}_{i_j} - \bar{\boldsymbol{a}})\right\|$. If we choose the constants C_1, C_2 appropriately, then for every good index i_j, U_{i_j} contains 0.95 fraction of the vertices in $V_{n_{i_j}} \cap Z$.

Lemma 31 then implies that when we throw away half of the sets $U'_1, ..., U'_m$ with the least Blue edge densities, then each of the remaining sets intersects some $V_{n_i} \cap Z$ in 0.9 fraction of the vertices.

Lemma 31 *There exists a constant $c > 0$ such that the following holds. Suppose we are given a set $X \subset Z$ of size $|X| = n/2k$. If for all $i \in 1, ..., k$*

$$|X \cap V_i| \leq 0.9|X|,$$

then with probability at least $1 - e^{-cn}$ *the number of Blue edges in the graph induced by* X *is at most* $an/16k - 0.09(a-b)n/16k$*. Conversely, if*

$$|X \cap V_i| \geq 0.95|X|$$

for some $i \in 1, ..., k$*, then with with probability at least* $1 - e^{-cn}$ *the number of Blue edges in the graph induced by* X *is at least* $an/16k - 0.09(a-b)n/16k$*.*

Proof

Let $e(X)$ denote the number of Blue edges in the graph induced by vertices in X. Suppose $|X \cap V_i| \leq 0.9|X|$ for all $i \in 1, ..., k$. Then

$$\boldsymbol{E}e(X) \leq an/16k - 0.09(a-b)n/8k.$$

To bound the probability that $\boldsymbol{E}e(X) \geq an/16k - .045(a-b)n/8k$, we can use Chernoff bound. Let $\delta = \frac{0.045(a-b)n/8k}{an/16k - 0.09(a-b)n/8k}$.

$$\mathbb{P}(\boldsymbol{E}e(X) \geq an/16k - .045(a-b)n/8k) \leq \exp\left(-\frac{(.045(a-b)n/8k)^2}{2an/16k + 0.045(a-b)n/8k}\right).$$

Similarly, suppose $|X \cap V_i| \geq 0.95|X|$ for some $i \in 1, ..., k$. Then

$$\boldsymbol{E}e(X) \geq an/16k - 0.05(a-b)n/16k.$$

To bound the probability that $\boldsymbol{E}e(X) \leq an/16k - .045(a-b)n/8k$, we can use Chernoff bound. Let $\delta = \frac{0.04(a-b)n/16k}{an/16k - 0.05(a-b)n/16k}$.

$$\mathbb{P}(\boldsymbol{E}e(X) \leq an/16k - .045(a-b)n/8k) \leq \exp\left(-\frac{(.04(a-b)n/16k)^2}{2an/16k + 0.04(a-b)n/16k}\right).$$

B.2. Proof of lemma 18

For notional convenience, let $U_i := Z \cap V_i$. We will use the following large deviation result (see page 36 in Boucheron et al. (2013) for e.g.) repeatedly

Lemma 32 *(Chernoff) If* X *is a sum of* n *iid indicator random variables with mean at most* $\rho \leq 1/2$*, then for any* $t > 0$

$$\max\{\mathbb{P}(X \geq \boldsymbol{E}X + t), \mathbb{P}(X \leq \boldsymbol{E}X - t)\} \leq \exp\left(-\frac{t^2}{2\,\boldsymbol{Var}X + t}\right) \leq \exp\left(-\frac{t^2}{2n\rho + t}\right).$$

By Theorem step1, there is a constant C such that if $\frac{(a-b)^2}{a+b} \geq C$ then by running **Spectral Partition** on the Red graph, we obtain, with probability $1 - o(1)$, sets $U_1', ..., U_k'$, where

$$|U_i' \backslash U_i| \leq 0.1n/2k.$$

In the rest, we condition on this event. The probability we will talk about in this section is based on the edges that go between vertices in Z.

Now we use the edges that go between vertices in Z. Consider $e = (u, v)$. If $u \in U_i, v \in U_j$ with $i \neq j$, then e is a Red edge with probability

$$\mu := b/2n.$$

Similarly, if $u, v \in U_i$, then e is a Red edge with probability

$$\tau := a/2n.$$

For any $u \in U_1$, the number of its neighbors in U'_j is at most

$$S_{1i}(u) := \sum_{i=1}^{.9n/2k} \xi_i^u + \sum_{j=1}^{.1n/2k} \zeta_j^u$$

Similarly, for any $u \in U_1$, the number of its neighbors in U'_1 is at least

$$S'_{11}(u) := \sum_{i=1}^{.9n/2k} \zeta_i^u + \sum_{j=1}^{.1n/2k} \xi_j^u.$$

After the correction sub-routine, if a vertex $u \in U_1$ is mislabeled then one of the following holds

- $S'_{1j} \geq \frac{a+b}{8}$ for some $j \neq 1$
- $S_{11} \leq \frac{a+b}{8}$.

By an application of Chernoff bound, probability that $S_{11} \leq \frac{a+b}{8}$ can be bounded by $\rho_1 = \exp(-0.04\frac{(a-b)^2}{k(a+b)})$. Similarly, for any fixed $j \neq 1$, $S'_{1j} \geq \frac{a+b}{8}$ is bounded by ρ_1. Therefore, the probability that any of these happens is bounded by $k\rho_1$. Therefore, number of vertices in U_1 that will be misclassified after the correction step is at most

$$M := \sum_{k=1}^{n/2k} \Gamma_k$$

where Γ_k are iid indicator random variables with mean ρ_1.

$$\boldsymbol{E}M \leq \frac{n}{2k} k \exp(-0.04\frac{(a-b)^2}{k(a+b)}).$$

Applying Chernoff's with $t := \frac{n}{2}\exp(-0.04\frac{(a-b)^2}{k(a+b)})$, we conclude that with probability $1 - o(1)$

$$M \leq \boldsymbol{E}M + t = n\exp(-0.04\frac{(a-b)^2}{k(a+b)}).$$

This implies that with probability $1 - o(1)$, number of mislabeled vertices in U_1 is

$$\leq n\exp(-0.04\frac{(a-b)^2}{k(a+b)}).$$

Set

$$\gamma := 2k\exp(-0.04\frac{(a-b)^2}{k(a+b)}).$$

Therefore, by a union bound over all i, we have that with probability $1-o(1)$ the output $U_1', U_2', ..., U_k'$ after the correction step form a γ-correct partition, with γ satisfying

$$\frac{(a-b)^2}{k(a+b)} = \frac{1}{0.04}\log\frac{2k}{\gamma} = 25\log\frac{2k}{\gamma},$$

proving our claim.

B.3. Proof of lemma 19

In this section, we show how we can merge $V_i \cap Y$ with $V_i \cap Z$ based on the Blue edges that go in between vertices in Y and Z. We can assume that that we are given a γ correct partition $U_1', .., U_k'$ of $U_1, ..., U_k$. Now we label the vertices in Y according to their degrees to U_i' as given in the **Merge** routine. Let us assume $\gamma \leq 0.1$. In the rest, we condition on this event, and the event that the maximum Red degree of a vertex is at most $\log^2 n$, which occurs with probability $1-o(1)$.

Now we use the Blue edges. Consider $e = (u, v)$. If e is not a red edge, and $u \in V_i \cap Y, v \in V_j \cap Z$, then e is a Blue edge with probability

$$\mu := \frac{b/2n}{1-\frac{b}{2n}}.$$

Similarly, if e is not a Red edge, and $u \in V_i \cap Z, v \in V_i \cap Z$, then e is a Blue edge with probability

$$\tau := \frac{a/2n}{1-\frac{a}{2n}}.$$

Thus, for any $u \in Y \cap V_i$, the number of Blue neighbors in U_j' is at most

$$S_j := \sum_{i=1}^{.9n/2k} \xi_i^u + \sum_{j=1}^{.1n/2k} \zeta_j^u$$

where ξ_i^u are iid indicator variables with mean μ and ζ_j^u are iid indicator variables with mean τ.

Similarly, for any $u \in Y \cap V_i$, the number of Blue neighbors in U_i' is at least

$$S_i' := \sum_{i=1}^{.9n/2k-d(u)} \zeta_i^u + \sum_{j=1}^{.1n/2k} \xi_j^u.$$

After the correction sub-routine, if a vertex u in $Y \cap V_i$ is misclassified then one of the following holds

- $S_j \geq \frac{a+b}{8k}$
- $S_i' \leq \frac{a+b}{8k}$.

Let ρ be the probability that at least one of the above events happens. Then the number of mislabeled vertices in the Y_2 is at most

$$M := \sum_{k=1}^{n/2} \Gamma_k$$

where Γ_k are iid indicator random variables with mean ρ. First we use Chernoff bound to estimate ρ. Consider

$$\rho_1 := \mathbb{P}\left(S_j \geq \frac{a+b}{8k}\right).$$

By definition, we have

$$\begin{aligned} \boldsymbol{E}S(u) &= \frac{0.9n\mu/k + 0.1n\tau/k}{2} \\ &= 0.9n(\frac{b/4kn}{1-\frac{b}{2n}}) + 0.1n(\frac{a/4kn}{1-\frac{a}{2n}}) \\ &= 0.9\frac{b}{4k} + 0.1\frac{a}{4k} + 0.9\frac{b}{4k}(\frac{1}{1-b/2n} - 1) + 0.1\frac{a}{4k}(\frac{1}{1-a/2n} - 1). \end{aligned} \tag{4}$$

Set

$$t := \frac{a+b}{8k} - \boldsymbol{E}S_j,$$

we have

$$t = 0.1\frac{a-b}{k} - 0.9\frac{b}{4k}(\frac{1}{1-b/2n} - 1) - 0.1\frac{a}{4k}(\frac{1}{1-a/2n} - 1) \geq 0.1\frac{a-b}{k} - 0.9\frac{b}{4k}\frac{b}{n} - 0.1\frac{a}{4k}\frac{a}{n} \geq 0.09\frac{a-b}{k},$$

for any sufficiently large n.

Applying Chernoff's bound, we obtain

$$\begin{aligned} \rho_1 &\leq \exp(-\frac{(0.09(a-b))^2}{k(0.9b/2 + .1a/2) + 0.09k(a-b)}) \\ &\leq \exp(-\frac{0.0324(a-b)^2}{k(a+b)}). \end{aligned}$$

By a similar argument, we get the same bound for

$$\rho_2 := \mathbb{P}\left(S'_i \leq \frac{a+b}{8k}\right).$$

Therefore, by a union bound, we have that

$$\rho \leq k\exp(-\frac{0.0324(a-b)^2}{k(a+b)}).$$

Thus, we can conclude that

$$\boldsymbol{E}M \leq \frac{n}{2}k\exp(-0.0324\frac{(a-b)^2}{k(a+b)}).$$

Applying Chernoff's with $t := \frac{n}{2}k\exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right)$, we conclude that with probability $1-o(1)$

$$M \leq \boldsymbol{E}M + t = nk\exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right).$$

This implies that with probability $1-o(1)$, the number of mislabeled vertices in Y is bounded by

$$nk\exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right).$$

Set

$$\gamma := 2k\exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right).$$

We have, with probability $1-o(1)$, γ correct partition of the vertices in Y, with γ satisfying

$$\frac{(a-b)^2}{k(a+b)} = \frac{1}{0.0324}\log\frac{2k}{\gamma} \leq 31\log\frac{2k}{\gamma},$$

proving our claim.

Appendix C. Censor Block Model

All we have to do now is to bound $\|E\|$. Let $\sigma^2 := p \geq \mathrm{Var}(\zeta_{i,j})$ for all (i,j). Y_0 is obtained by zeroing out rows and columns of Y of high degree. We then have the following lemma. The proof is essentially the same as corollary 13, so we skip the details.

Lemma 33 $0 < \epsilon_0 \leq \epsilon < \frac{1}{2}$. *Then there exist constants* C, C_1 *such that if* $p \geq \frac{C}{n}$, *then with probability* $1-o(1)$, $\left\|Y_0 - \bar{Y}\right\| \leq C_1\sigma\sqrt{n} = C_1\sqrt{np}$.

Since the second eigenvalue of $\bar{Y}$ is $p(1-2\epsilon)n$, to make the angle between the eigenspace spanned by the two eigenvectors corresponding to the top two eigenvalues small, we need to assume

$$\frac{p(1-2\epsilon)n}{\sqrt{np}}$$

is sufficiently large. The assumption

$$np \geq \frac{C_2}{(1-2\epsilon)^2}$$

in theorem 6 is precisely this.

Appendix D. Proof of Lemma 21

This proof is essentially same as that in Feige and Ofek (2005). Let us first define the following sets. For $\gamma_k := 2^k$,

$$S_k := \left\{ i : \frac{\gamma_{k-1}}{\sqrt{n}} < x_i \leq \frac{\gamma_k}{\sqrt{n}} \right\}, s_k := |S_k|, k = \lfloor \log \frac{\sqrt{d}}{n} \rfloor, .., 0, 1, 2, ..., \lceil \log \sqrt{n} \rceil$$

and

$$T_k := \left\{ i : \frac{\gamma_{k-1}}{\sqrt{n}} < y_i \leq \frac{\gamma_k}{\sqrt{n}} \right\}, t_k := |T_k|, k = \lfloor \log \frac{\sqrt{d}}{n} \rfloor, .., 0, 1, 2, ..., \lceil \log \sqrt{n} \rceil.$$

Further, we use the notation $\mu_{i,j} := s_i t_j \frac{d}{n}$ and $\lambda_{i,j} := e(S_i, T_j)/\mu_{i,j}$. We then have

$$\begin{aligned}
\sum_H x_i \widetilde{A}_{i,j} y_j &\leq \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} s_i t_j \frac{d}{n} \lambda_{i,j} \frac{\gamma_i}{\sqrt{n}} \frac{\gamma_j}{\sqrt{n}} \\
&= \sqrt{d} \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} s_i \frac{\gamma_i^2}{n} t_j \frac{\gamma_j^2}{n} \frac{\lambda_{i,j}\sqrt{d}}{\gamma_i\gamma_j} \\
&= \sqrt{d} \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i \beta_j \sigma_{i,j}.
\end{aligned}$$

In the last line, we have used the following notation $\alpha_i := s_i \frac{\gamma_i^2}{n}, \beta_j := t_j \frac{\gamma_j^2}{n}, \sigma_{i,j} := \frac{\lambda_{i,j}\sqrt{d}}{\gamma_i\gamma_j}$. In this notation, we can write 2 as follows:

$$\sigma_{i,j}\alpha_i \log \lambda_{i,j} \leq c_3 \frac{\gamma_i}{\gamma_j\sqrt{d}} \left[2\log\gamma_j + \log\frac{1}{\beta_j} \right]. \tag{5}$$

Now we bound $\sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j\sigma_{i,j}$ by a constant. We note that $\sum_i \alpha_i \leq 4$ and $\sum_i \beta_i \leq 4$. We now consider 6 cases.

1. $\sigma_{i,j} \leq 1$:

$$\begin{aligned}
\sqrt{d} \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j\sigma_{i,j} &\leq \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j \\
&\leq \sqrt{d}(\sum_i \alpha_i)(\sum_i \beta_i) \\
&\leq 16\sqrt{d}.
\end{aligned}$$

2. $\lambda_{ij} \leq c_2$: Since $\gamma_i\gamma_j \geq \sqrt{d}$ we have in this case $\sigma_{i,j} \leq c_2$. Therefore,

$$\begin{aligned}
\sqrt{d} \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j\sigma_{i,j} &\leq \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j c_2 \\
&\leq c_2\sqrt{d}(\sum_i \alpha_i)(\sum_i \beta_i) \\
&\leq 16c_2\sqrt{d}.
\end{aligned}$$

3. $\gamma_i > \sqrt{d}\gamma_j$: Since the maximum degree is $\leq c_1 d$, we have that $\lambda_{i,j} \leq c_1 n/t_j$. Therefore,

$$\begin{aligned}
\sqrt{d} \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j\sigma_{i,j} &= \sqrt{d}\sum_i \left(\alpha_i \sum_{j:\gamma_i\gamma_j\geq\sqrt{d}} \beta_j \frac{\lambda_{i,j}\sqrt{d}}{\gamma_i\gamma_j}\right) \\
&\leq \sqrt{d}\sum_i \left(\alpha_i \sum_{j:\gamma_i\gamma_j\geq\sqrt{d}} b_j \frac{\gamma_j^2}{n}\frac{(c_1 n/b_j)\sqrt{d}}{\gamma_i\gamma_j}\right) \\
&= \sqrt{d}\sum_i \left(\alpha_i \sum_{j:\gamma_i\gamma_j\geq\sqrt{d}} c_1\sqrt{d}\frac{\gamma_j}{\gamma_i}\right) \\
&\leq \sqrt{d}\sum_i (\alpha_i c_1 \times 2) \\
&\leq 2c_1\sqrt{d}\sum_i \alpha_i \\
&\leq 8c_1\sqrt{d}.
\end{aligned}$$

4. We now assume that we are **not** in cases $1-3$. Therefore, we can assume that 5 holds. We consider the following sub cases.

   (a) $\log\lambda_{i,j} > (1/4)[2\log\gamma_j + \log(1/\beta_j)]$: 5 implies that $\sigma_{i,j}\alpha_i \leq 4c_3(\gamma_i/\gamma_j\sqrt{d})$. Therefore,

$$\begin{aligned}
\sqrt{d} \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j\sigma_{i,j} &= \sqrt{d}\sum_j \beta_j \sum_{i:\gamma_i\gamma_j\geq\sqrt{d}} \alpha_i\sigma_{i,j} \\
&\leq \sqrt{d}\sum_j \beta_j \sum_{i:\gamma_i\gamma_j\geq\sqrt{d}} 4c_3\frac{\gamma_i}{\gamma_j\sqrt{d}} \\
&\leq \sqrt{d}\sum_j \beta_j \times 8c_3 \\
&\leq 32c_3\sqrt{d}.
\end{aligned}$$

   Above we made use of the fact that we are not in case 3, and that $\sum_{i:\gamma_i\gamma_j\geq\sqrt{d}} 4c_3\frac{\gamma_i}{\gamma_j\sqrt{d}}$ is a geometric sum.

   (b) $2\log\gamma_j \geq \log(1/\beta_j)$: We can assume we are not in case (a), and hence $\lambda_{i,j} \leq \gamma_j$. Combined with the fact that we are not in case 1, we have that $\gamma_i \leq \sqrt{d}$. Since we are not in case 2, we can assume that $\log\lambda_{i,j} \geq 1$ and hence $\sigma_{i,j}\alpha_i \leq c_3\frac{\gamma_i}{\gamma_j\sqrt{d}}4\log\gamma_j$. Therefore,

$$\begin{aligned}
\sqrt{d} \sum_{i,j:\gamma_i\gamma_j \geq \sqrt{d}} \alpha_i\beta_j\sigma_{i,j} &= \sqrt{d}\sum_j \left(\beta_j \sum_{i:\gamma_i\gamma_j\geq\sqrt{d}} \alpha_i\frac{\lambda_{i,j}\sqrt{d}}{\gamma_i\gamma_j}\right) \\
&\leq \sqrt{d}\sum_j \left(\beta_j \sum_{i:\gamma_i\gamma_j\geq\sqrt{d}} 4c_3\frac{\gamma_i}{\sqrt{d}\gamma_j}\log\gamma_j\right)
\end{aligned}$$

$$\leq \sqrt{d}\left(\sum_{j} \beta_j 4c_3 \sum_{i:\gamma_i\gamma_j \geq \sqrt{d}} \frac{\gamma_i}{\sqrt{d}}\right)$$
$$\leq \sqrt{d}\sum_{j} \beta_j 4c_3 \times 2$$
$$\leq 32c_3\sqrt{d}.$$

(c) $2\log\gamma_j \leq \log 1/\beta_j$: Since we are not in (a) we have $\log\lambda_{i,j} \leq \log\frac{1}{\beta_j}$. It follows that

$$\sigma_{i,j} = \frac{\lambda_{i,j}\sqrt{d}}{\gamma_i\gamma_j} \leq \frac{1}{\beta_j}\frac{\sqrt{d}}{\gamma_i\gamma_j}.$$

Therefore:

$$\sqrt{d}\sum_{i,j:\gamma_i\gamma_j\geq\sqrt{d}} \alpha_i\beta_j\sigma_{i,j} = \sqrt{d}\sum_{i}\left(\alpha_i\sum_{j:\gamma_i\gamma_j\geq\sqrt{d}} \beta_j\sigma_{i,j}\right)$$
$$= \sqrt{d}\sum_{i}\left(\alpha_i\sum_{j:\gamma_i\gamma_j\geq\sqrt{d}} \beta_j\frac{\lambda_{i,j}\sqrt{d}}{\gamma_i\gamma_j}\right)$$

$$
\begin{aligned}
&\leq \sqrt{d} \sum_i \left(\alpha_i \sum_{j:\gamma_i\gamma_j \geq \sqrt{d}} \frac{\sqrt{d}}{\gamma_i\gamma_j} \right) \\
&\leq \sqrt{d} \sum_i (\alpha_i \times 2) \\
&\leq 8\sqrt{d}.
\end{aligned}
$$

■