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

Peter Chin ∗ Boston University spchin@cs.bu.edu
Anup Rao † Yale University anup.rao@yale.edu
Van Vu ‡ Yale University van.vu@yale.edu

January 22, 2015

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.

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 $a/n$ and an edge between vertices across different blocks appear with probability $b/n$, where $a > b > 0$. Given an instance of this graph, we would like to identify the two blocks as correctly as possible. We will consider the case of more than 2 blocks later in the paper.

This 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. the 1980s [BCLS87]; see [DF89] [Rav87] [JS93] [McS01] [Vu14] 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. Improving an earlier result in [McS01], Vu [Vu14] shows that one can efficiently obtain a complete recovery if $\frac{(a-b)^2}{a+b} \geq C \log n$ and $a, b \geq C \log n$ for some sufficiently large constant $C$.

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 [Bol01]. Apparently, there is no way to tell these 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 1.1 A partition $|V'_1 \cup V'_2|$ of $V_1 \cup V_2$ is $\gamma$-correct if $|V_i \cap V'_i| \geq (1-\gamma)n$.

In [CO10], Coja-Oglan proved

∗ NSF-DMS-1222567 and by Air Force Office of Scientific Research (FA9550-12-1-0136).
† This work was partially supported by NSF grant CCF-1111257.
‡ V. H. Vu is supported by research grants DMS-0901216 and AFOSAR-FA-9550-09-1-0167 and a grant from Draper Lab.
**Theorem 1.2** 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 \( \gamma \)-correct partition using a polynomial time algorithm.

Coja-Oglan proved Theorem 1.2 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. (We would like to thank E. Abbe for communicating this conjecture.) Even when the log term is removed, an important question is to find out the optimal relation between the accuracy \( \gamma \) and the ratio \( \frac{(a-b)^2}{a+b} \). This is the main goal of this paper.

**Theorem 1.3** 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 \( \gamma \)-correct partition with probability \( 1 - o(1) \) using a simple spectral algorithm.

The constants \( C_1, C_2 \) 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 3.10 yields the following corollary

**Corollary 1.4** There are constants \( C_0 \) and \( \epsilon \) 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 \( \gamma \)-correct partition with probability \( 1 - o(1) \) using a simple spectral algorithm.

In parallel to our study, Zhang and Zhou [ZZ], proving a minimax rate result, showed that there is a constant \( c > 0 \)

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

then one cannot recover a \( \gamma \)-correct partition (in expectation), regardless the algorithm. Furthermore, for some other constant \( C \), if

\[
\frac{(a-b)^2}{a+b} \geq C \log \frac{1}{\gamma},
\]

then there is an algorithm which recovers a \( \gamma \)-correct partition (in expectation). As customary in minimax rate analysis, algorithms considered by Zhang and Zhou are allowed to have arbitrary complexity.

Our Theorem 3.10 makes the above theoretical bound effective. We design a fast algorithm which obtains a \( \gamma \)-correct partition under the optimal condition \( \frac{(a-b)^2}{a+b} \geq C \log \frac{1}{\gamma} \). As a matter of fact, our algorithm guarantees \( \gamma \)-correctness with high probability, instead of in expectation.

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 \( |V_i| = n \) with edge probabilities \( a_n \) between vertices within the same block and \( b_n \) between vertices in different blocks.

**Theorem 1.5** For any \( \epsilon, \gamma > 0 \) and a fixed integer \( k > 0 \), there exists constants \( C_1, C_2 \) such that if

1. \( a > b \geq C_1 \)
2. \(a - b \geq C_2\sqrt{a + b}\), and
3. \(a = o\left(\frac{n}{\log n}\right)\),
then we can find a \(\gamma\)-correct partition with probability at least \(1 - \epsilon\) using a simple spectral algorithm.

Similarly to the \(k = 2\) case, one can have \(C_2 = C_3 \log \frac{1}{\gamma}\), where \(C_3\) is a constant not depending on \(\gamma\). Furthermore, one can choose \(C_1 = C_4 \frac{1}{\sqrt{\epsilon}}\), where \(C_4\) is a constant not depending on \(\epsilon\).

To conclude this section, let us present an application of our method to the Censor Block Model studied by Abbe et. al. in [ABBS14]. 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 \(x = (x_1, \ldots, x_{2n})\) be the indicator vector of \(V_2\). Let \(z\) be a random noise vector whose coordinates \(z_{e_i}\) are i.i.d Bernoulli(\(\epsilon\)) (taking value 1 with probability \(\epsilon\) and 0 otherwise), where \(e_i\) are the edges of \(G\).

Given a noisy observation
\[
y = B_G x \oplus z
\]
where \(\oplus\) is the addition in mod 2, one would like identify the blocks. In [ABBS14], the authors proved that exact recovery (\(\gamma = 0\)) is possible if and only if \(\frac{np}{\log n} \geq \frac{2}{(1 - 2\epsilon)^2} + o\left(\frac{1}{(1 - 2\epsilon)^2}\right)\) 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 1.6** 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 \(\gamma\)-correct partition with probability \(1 - o(1)\), using a simple spectral algorithm.

Let us also mention a related, interesting, problem, where the purpose is to do better than a random guess (in our term, to find a partition which is \((1/2 - \epsilon)\)-correct). It was conjectured in [DKMZ11] that this is possible if and only if \((a - b)^2 > (a + b)\). This conjecture has been settled recently by Massoulie [MNS12][Mas13] and Mossel et. al. [MNS13].

*Added in proof.* In a recent discussion, Zhou pointed out that Mossel et. al. also improved (theorem 2.3 in [MNS13]) Coja-Oghlan’s result by removing the \(\log (a + b)\) term. Their result, however, does not specify any relation between \(\gamma\) and \(C\). Furthermore, their algorithm is completely different from ours.

The rest of the paper is organized as follows. In section 2, we describe our algorithm for Theorem 2.1 and an overview of the proof. The full proof comes in sections 3 and 4. In section 5, we show how the techniques can be extended to the \(k\) block case and prove theorem 1.5. Finally, in section 6, we prove theorem 1.6.

## 2 Our algorithm

Our algorithm will have two steps. First we use a spectral algorithm to recover a partition where is dependence between \(\gamma\) 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 2.1. Let \(\tilde{A}_0 = \mathbb{E}A_0\) and \(E_0 = A_0 - \tilde{A}_0\). Then \(\tilde{A}_0\) is a matrix with rank two with the two non zero eigenvalues \(\lambda_1 = a + b\) and \(\lambda_2 = a - b\). The eigenvector \(u_1\) corresponding to the eigenvalue \(a + b\) has coordinates
\[
u_1(i) = \frac{1}{\sqrt{2n}} \text{ for all } i \in V
\]
and eigenvector \(u_2\) corresponding to the eigenvalue \(a - b\) has coordinates
\[
u_2(i) = \begin{cases} \frac{1}{\sqrt{2n}} & \text{if } i \in V_1 \\ -\frac{1}{\sqrt{2n}} & \text{if } i \in V_2. \end{cases}
\]
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 $v_1$, the projection of all-ones vector on to $W$.

5. Let $v_2$ be the unit vector in $W$ perpendicular to $v_1$.

6. Sort the vertices according to their values in $v_2$, 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 us provide the reader with the intuition behind this algorithm. Notice that the second eigenvector of $\bar{A}_0$ identifies the partition. We would like to use the second eigenvector of $A_0$ to approximate it. 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.

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

$$A = \bar{A} + E = \bar{A}_0 + \Delta + E.$$

The key observation is that $\|E\|$ is significantly smaller than $\|E_0\|$. In the next section we will show that $\|E\| = O(\sqrt{d})$, with probability $1 - o(1)$, while $\|E_0\|$ is $\Theta(\sqrt{\log n})$, with probability $1 - o(1)$. Furthermore, we could show that $\|\Delta\|$ is only $O(1)$ 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 2.1** There are constants $C_0$ and $C_1$ such that the following holds. For any constants $a > b > C_0$ and $\gamma > 0$ satisfying $\frac{(a-b)^2}{a+b} \geq C_1 \frac{1}{\gamma}$, then with probability $1 - o(1)$, Spectral Partition outputs a $\gamma$-correct partition.

**Remark 2.2** The parameter $d := a + b$ can be estimated very efficiently from the adjacency matrix $A$. We take this as input for a simpler exposition.
**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**

Step 2 is a further correction that gives us the optimal (logarithmic) dependence between $\gamma$ 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 $|V_1 \setminus V_1'| \leq \frac{1}{10}$, 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

$$0.1b + 0.9a > \frac{a + b}{2} > 0.9b + 0.1a$$

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 dependence. 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.

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.
3 First step: Proof of Theorem 2.1

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 $ar{W}$ 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 second eigenvector**: If $\sin \angle(W, \bar{W})$ is small, we find a vector $v_2 \in W$ such that $\sin \angle(v_2, u_2)$ is small.

To do the first part, recall that $A = \bar{A}_0 + \Delta + E$; we first prove that $\|\Delta\|$ and $\|E\|$ are small with probability $1 - o(1)$.

### 3.1 Bounding $\|\Delta\|$  

This is the easier part, as it will be sufficient to bound the number of vertices of high degrees. We need the following

**Lemma 3.1** There exist a constant $d_0$ such that if $d := a + b \geq d_0$, then with probability $1 - \exp(-\Omega(a^{-2}n))$ not more than $a^{-3}n$ vertices have degree $\geq 20d$. 

---

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.
If there are at most \(a^{-3}n\) vertices with degree \(\geq 20d\), then by definition, \(\Delta\) 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 \(\|\Delta\|_{HS} \leq \sqrt{2a^{-1/2}}\).

**Corollary 3.2** For \(d_0\) sufficiently large, with probability \(1 - \exp(-\Omega(a^{-3}n))\), \(\|\Delta\| \leq \sqrt{2a^{-1/2}} \leq 1\).

One can prove Lemma 3.1 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 \(\mu_{E(X)} := E(|E(X)|)\) satisfies

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

Let \(\delta_1 := \frac{2}{c} \geq \frac{2cnd}{\mu_{E(X)}}\), then Chernoff bound (see [AS04] for example) gives

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

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 \(\delta_2 := 4 \geq \frac{8cnd}{\mu_{E(X,X_c)}}\), then by Chernoff bound

\[
\mathbb{P}(|E(X,X_c)| \geq 8cnd) \leq \left(\frac{\exp(\delta_2 - 1)}{\delta_2^{\delta_2}}\right)^{\mu_{E(X,X_c)}} \leq \exp (-c(2-c)an).
\]

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

\[
\mathbb{P}(|E(X)| \geq 2cnd) \leq \exp (-0.75 \log(a)a^{-2}n)\]

\[
\mathbb{P}(|E(X,X_c)| \geq 8cnd) \leq \exp (-a^{-2}n).
\]

There are at most

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

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.

### 3.2 Norm of sparse random matrices

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

**Lemma 3.3** Suppose \( M \) is random symmetric matrix with zero on the diagonal whose entries above the diagonal are independent with the following distribution

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

Let \( \sigma \) 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 \( 20\sigma^2n \) positive entries. Then with probability \( 1 - o(1) \), \( \|M_1\| \leq C\sigma\sqrt{n} \) for some constant \( C > 0 \).

We start by proving a simpler result.

**Lemma 3.4** Let \( M \) be random symmetric matrix of size \( n \) with zero diagonal whose entries above the diagonal are independent with the following distribution

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

Let \( \sigma^2 \geq C_1\log \frac{n}{n} \) 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 3.4. A weaker bound \( C\sigma\sqrt{n}\log n \) follows easily from Alshwede-Winter type matrix concentration results (see [Tro12]). To prove the claimed bound, we need to be more careful and follow the \( \epsilon \)-net approach by Kahn and Szemerédi for random regular graphs in [FKS89] (see also [AK94, FO05, ?]).

Consider a \( \frac{1}{2} \)-net \( \mathcal{N} \) of the unit sphere \( S^n \). We can assume \( |\mathcal{N}| \leq 5^n \). It suffices to prove that there exists a constant \( C'_2 \) such that with probability \( 1 - o(1) \), \( \|x^TMy\| \leq C'_2\sigma\sqrt{n} \) for all \( x, y \in \mathcal{N} \).

For two vectors \( x, y \in \mathcal{N} \), we follow an argument of Kahn and Szemerédi [FKS89] and call all pairs \((i,j)\) such that \( |x_iy_j| \leq \frac{\sigma}{\sqrt{n}} \) light and all remaining pairs heavy and denote these two classes by \( L \) and \( H \) respectively. We have

\[
x^TMy = \sum_{i,j} x_iM_{ij}y_j = \sum_L x_iM_{i,j}y_j + \sum_H x_iM_{i,j}y_j.
\]

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 \( X := \sum_L x_iM_{i,j}y_j \) as \( \sum_{(i,j) \in L, i \geq j} M_{i,j}a_{i,j} \), where

\[
a_{i,j} = \begin{cases} 
x_iy_j + x_jy_i & \text{if } (i, j), (j, i) \in L \\
x_iy_j & \text{if } (i, j) \in L \\
x_jy_i & \text{if } (j, i) \in L
\end{cases}
\]

By the definition light pairs, \( |a_{i,j}| \leq 2\frac{\sigma}{\sqrt{n}} \). Also, since \( x \) and \( y \) are unit vectors, \( \sum_{i,j} a_{i,j}^2 \leq 4 \). Therefore, by Bernstein’s bound (see page 36 in [BLM13] for e.g.)
\[ \mathbb{P}(X > t) \leq \exp \left( \frac{-\frac{1}{2} t^2}{4 \sigma^2 + \frac{1}{2} \frac{\sigma}{\sqrt{n}} t} \right). \]

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 \( 1 - \exp(-3n) \), \( |\sum_\mathcal{H} x_i M_{i,j} y_j| \leq 10 \sigma \).

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 \( \sum_{H} |x_i y_j| \leq \sqrt{n} \). Let \( A_{i,j} := M_{i,j} + p_{i,j} \), then

\[
\sum_{H} x_i M_{i,j} y_j = \sum_{H} x_i A_{i,j} y_j - \sum_{H} p_{i,j} x_i y_j.
\]

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

Lemma 3.5 Let \( \tilde{G} = (\tilde{V}, \tilde{E}) \) be any graph whose adjacency matrix is denoted by \( \tilde{A} \), and \( x, y \) be any two unit vectors. Let \( \tilde{d} \) be such that the maximum degree \( \leq c_1 \tilde{d} \). Further, let \( \tilde{d} \) satisfy the property that for any two subsets of vertices \( S, T \subset \tilde{V} \) one of the following holds for some constants \( c_2 \) and \( c_3 \):

\[
\frac{e(S, T)}{|S||T| \tilde{d}^2 / n} \leq c_2 \quad (3.1)
\]

\[
e(S, T) \log \left( \frac{e(S, T)}{|S||T| \tilde{d}^2 / n} \right) \leq c_3 |T| \log \frac{n}{|T|} \quad (3.2)
\]

then \( \sum_{H} x_i A_{i,j} y_j \leq \max(16, 8c_1, 32c_2, 32c_3) \sqrt{\tilde{d}}. \) Here \( H := \{(i, j)||x_i y_j| \geq \sqrt{\tilde{d}} / n \} \).

The proof appears in appendix.

Lemma 3.6 Let \( \tilde{d} := \sigma^2 n \). Then with probability \( 1 - o(1) \), the maximum degree in the graph \( G_A \) is \( \leq 20 \tilde{d} \) with probability \( 1 - o(1) \) and for any \( S, T \subset V \) one of the conditions \( (\ref{eq:cond1}) \) or \( (\ref{eq:cond2}) \) holds.

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

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

\[
A_{i,j} = \begin{cases} 1 & \text{w.p. } p_{i,j} \\ 0 & \text{w.p. } 1 - p_{i,j} \end{cases}.
\]

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

\[
\mu = \mathbb{E} X = \sum_i p_{ik} \leq \sigma^2 n.
\]

For any \( l \geq 4 \), Chernoff bound (see [AS04]) implies that
\[
\mathbb{P}(X > l\sigma^2n) \leq \exp\left(-\frac{\sigma^2nl\ln l}{3}\right) \\
\leq \exp\left(-\frac{l\log n}{3}\right).
\]

Applying this with \(l = 20\), and taking a union bound over all the vertices, we can bound the maximum degree by \(20\sigma^2n\). 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 \(\mathbf{E}X \leq \sigma^2|S||T|\). If \(|T| \geq \frac{n}{e}\), then since the maximum degree is \(\leq 20\sigma^2n\), we have \(e(S,T) \leq |S|20\sigma^2n \leq 20e\sigma^2|S||T|\), giving us ?? 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 [FO05], if we choose \(l = \max(l', 4)\), we can bound the above probability by \(\exp\left(-\frac{l\ln(l)\sigma^2|S||T|}{3}\right)\left(\frac{n}{|S|}\right)^{|T|} \leq \frac{1}{n^3}\). 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 ?? or ?? holds with probability \(1 - o(1)\).

**Proof of Lemma 3.3.** Now we are ready to prove Lemma 3.3 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 3.4 implies that with probability at least \(1 - \exp(-3n)\), for all \(x, y \in \mathcal{N}_{1/2}\), \(|\sum_L x_i(M_0)_{ij}y_j| \leq 10\sigma\sqrt{n}\). 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 3.5 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 (??) or (??) 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 (??) holds for \(A_1\) since the degrees are bounded.

Now, to bound the norm of matrix \(E\), we just appeal to 3.3. Suppose \(a > b \geq C_0\), for a large enough constant \(C_0\) to be determined later. Since \(A = A_0 + \Delta + E\) and we have bounded \(\Delta\), it remains to bound \(\|E\|\). Note that
\[
(E_0)_{ij} = \begin{cases} 
1 - \frac{a}{n} & \text{w.p.} \frac{a}{n} \\
-\frac{a}{n} & \text{w.p.} \frac{a}{n} \\
1 - \frac{a}{n} & \text{w.p.} 1 - \frac{a}{n}
\end{cases}
\]
if \( i, j \) belongs to the same community and

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

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

\[
\text{Var}((E_0)_{ij}) \leq \frac{a}{n} (1 - \frac{a}{n}) \leq \frac{d}{n}.
\]

Lemma 3.3 implies

**Corollary 3.7** 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 \( 1 - o(1) \).

### 3.3 Bounding the angle

Now, we bound the angle \( \angle(\bar{W}, W) \). Let us fix the following additional notation in this section. Let \( \bar{v}_1, \bar{v}_2 \) be eigenvectors of \( \bar{A}_0 \) corresponding to the largest two eigenvalues \( \lambda_1 \geq \lambda_2 \). \( v_1, v_2 \) be eigenvectors of \( A = A_0 + \Delta + E \) corresponding to the largest two eigenvalues \( \lambda_1' \geq \lambda_2' \). Further, \( \bar{W} := \text{Span}\{\bar{v}_1, \bar{v}_2\} \) and \( W := \text{Span}\{v_1, v_2\} \).

**Lemma 3.8** For any constant \( \gamma < 1 \), we can choose constants \( C_2 \) and \( C_3 \) such that such that if

1. \( a - b \geq C_2\sqrt{a + b} = C_2\sqrt{d} \)
2. \( b \geq C_3 \)

then, \( \sin(\angle \bar{W}, W) \leq c < 1 \) with probability \( 1 - o(1) \).

**Proof of Lemma 3.8:** Let \( C_3' \) be a constant such that if \( a > b \geq C_3' \), then theorem 3.2 holds giving us \( \|\Delta\| \leq 1 \). From lemma 3.7 we have that \( \|E\| \leq C\sqrt{d} \). Now, \( \lambda_2' \geq \lambda_2 - \|E\| \geq a - b - C\sqrt{a + b} \). For any constant \( C_1 \), there exists constants \( C_2 \) and \( C_3'' \) such that if \( a - b \geq C_2\sqrt{a + b} \) and \( b \geq C_3'' \), we can ensure that \( \lambda_2' \geq C_1\sqrt{a + b} \). The lemma then follows from the Davis-Kahan [Dav63] [Bha97] bound for matrices \( \bar{A}_0 \) and \( A \), which gives \( \sin(\angle W, \bar{W}) \leq \frac{\|E + \Delta\|}{\lambda_2'} \). Let \( C_3 = \max(C_3', C_3'') \). Therefore, the lemma follows by choosing \( C_1 \) big enough.

### 3.4 Recovery

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

**Lemma 3.9** If \( \sin(\angle \bar{W}, W) \leq c \leq \frac{1}{4} \), then we can find a vector \( v \in W \) such that \( \sin(\angle v, \bar{v}_2) \leq 2\sqrt{c} \).

**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 \( \bar{v}_1 \) 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 $\mathbf{v}_1, \mathbf{v}_2 \in \tilde{W}$, we have that $\|P_W \mathbf{v}_i - \mathbf{v}_i\|_2 \leq c$ for $i = 1, 2$. Let us define $\mathbf{u}_i := P_W \mathbf{v}_i$ and $\mathbf{x}_i := \mathbf{u}_i - \mathbf{v}_i$ (note that $\|\mathbf{x}_i\| \leq c$) for $i = 1, 2$. We will now show that the vector $\mathbf{v} \in W$ perpendicular to $\mathbf{u}_1$ is close to $\mathbf{v}_2$. Let $\mathbf{u}_1 = \mathbf{u}_2 - \frac{\mathbf{u}_1^T \mathbf{u}_2}{\|\mathbf{u}_1\|^2} \mathbf{u}_1$, it is then clear that $\|\mathbf{u}_1\| \leq 1$. Note that $|\mathbf{u}_1^T \mathbf{u}_2| = |\mathbf{v}_1^T \mathbf{x}_2 + \mathbf{v}_2^T \mathbf{x}_1 + \mathbf{x}_1^T \mathbf{x}_2| \leq 2c + c^2$. We have,

$$u_1^T \mathbf{v}_2 = u_2^T \mathbf{v}_2 - \frac{(u_1^T \mathbf{u}_2)(v_2^T \mathbf{u}_1)}{\|\mathbf{u}_1\|^2}$$

$$|u_1^T \mathbf{v}_2| \geq 1 - c - \frac{(2c + c^2)c}{(1-c)^2}$$

$$\geq 1 - 2c.$$

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

$$|\mathbf{v}^T \mathbf{v}_2| \geq |\mathbf{u}_1^T \mathbf{v}_2| \geq 1 - 2c.$$

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

Lemmas 3.8 and 3.9 together give

**Corollary 3.10** For any constant $c < 1$, we can choose constants $C_2$ and $C_3$ in lemma 3.8 and find a vector $\mathbf{v}$ such that $\sin(\angle \mathbf{v}_2, \mathbf{v}) \leq c < 1$ with probability $1 - o(1)$.

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

**Lemma 3.11** If $\sin(\angle \mathbf{v}_2, \mathbf{v}) < c \leq 0.5$, then we can identify at least a $(1 - \frac{2}{\sqrt{3}}c)$ fraction of vertices from each block correctly.

**Proof of Lemma 3.11:** Let us define two sets of vertices, $V'_1 = \{i | \mathbf{v}(i) > 0\}$ and $V'_2 = \{i | \mathbf{v}(i) < 0\}$. One of the sets will have less than or equal to $\frac{c}{2}$ vertices, let us assume without loss of generality that $|V'_1| \leq \frac{c}{2}$. Writing $\mathbf{v} = c_1 \mathbf{v}_2 + \mathbf{err}$, for a vector $\mathbf{err}$ perpendicular to $\mathbf{v}_2$ and $\|\mathbf{err}\| < c$. We also have $c_1 > \sqrt{1 - c^2}$. Since $\|\mathbf{err}\| < c$, not more than $\frac{c}{\sqrt{1 - c^2}} \cdot n$ coordinates of $\mathbf{err}$ can be bigger than $\frac{1-c^2}{\sqrt{n}}$. Since $\mathbf{v} = c_1 \mathbf{v}_2 + \mathbf{err}$ at least $1 - \frac{c}{\sqrt{1 - c^2}} > 1 - \frac{c}{\sqrt{3}}$ (since $c \leq 0.5$) fraction of vertices with $\mathbf{v}_2(i) = \frac{1}{\sqrt{n}}$ will have $\mathbf{v}(i) > 0$. Therefore, we get that there are at least $(1 - \frac{2}{\sqrt{3}}c)n$ vertices belonging to the first block.

**4 Second step: Analysis of Correction**

We will use the following large deviation result repeatedly

**Lemma 4.1** (Chernoff) If $X$ is a sum of $n$ iid indicator random variables with mean at most $\rho \leq 1/2$, then for any $t > 0$

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

In the Red graph, the edge densities are $a/2n$ and $b/2n$, respectively. By Theorem 2.1, 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
\[ |V_i \setminus V'_i| \leq .1n. \]

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 \( \xi_i^u \) are iid indicator variables with mean \( \mu \) and \( \zeta_j^u \) are iid indicator variables with mean \( \tau \).

Similarly, for any \( u \in V'_i \cap V_2 \), the number of its Blue neighbors in \( V'_2 \) is at least

\[ S'(u) := \sum_{i=1}^{.9n-d(u)} \xi_i^u + \sum_{j=1}^{1n} \zeta_j^u, \]

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

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

- \( u \in V'_i \cap V_1 \) and \( S_u \geq \frac{a+b}{4} \).
- \( u \in V'_i \cap V_2 \) and \( S'_u \leq \frac{a+b}{4} \).

Let \( \rho_1, \rho_2 \) be the probability of the above events. Then the number of misclassified vertices in the (corrected) set \( V'_i \) is at most

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

where \( \Gamma_k \) are iid indicator random variables with mean \( \rho_1 \) and \( \Lambda_l \) are iid indicator random variables with mean \( \rho_2 \).

The rest is a simple computation. First we use Chernoff bound to estimate \( \rho_1, \rho_2 \). Consider

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

By definition, we have

\[ \mathbb{E}S(u) = 0.9n\mu + 0.1n\tau = 0.9n\left( \frac{b/2n}{1 - \frac{b}{2n}} \right) + 0.1n\left( \frac{a/2n}{1 - \frac{a}{2n}} \right) = 0.9 \frac{b}{2} + 0.1 \frac{a}{2} + 0.9 \frac{b}{2} \left( \frac{1}{1 - b/2n} - 1 \right) + 0.1 \frac{a}{2} \left( \frac{1}{1 - a/2n} - 1 \right). \]

(4.3)
Set 
\[ t := \frac{a + b}{4} - ES(u), \]
we have

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

for any sufficiently large \( n \).

Applying Chernoff’s bound, we obtain
\[ \rho_1 \leq \exp\left( -\frac{(0.19(a - b))^2}{2(0.9n\mu + 1.0\tau) + 0.19(a - b)} \right). \]

By (??), one can show that \( 2(0.9n\mu + 1.0\tau) + 0.19(a - b) = 0.71b + 0.29a + o(1) \leq \frac{a + b}{2} \). It follows that
\[ \rho_1 \leq \exp\left( -0.072 \frac{(a - b)^2}{a + b} \right). \]

By a similar argument, we obtain the same estimate for \( \rho_2 \) (the contribution of the term \( d(u) \leq \log^2 n \) is negligible). Thus, we can conclude that
\[ EM \leq 1.1n \exp\left( -0.072 \frac{(a - b)^2}{a + b} \right). \]

Applying Chernoff’s with \( t := 0.9n \exp\left( -0.072 \frac{(a - b)^2}{a + b} \right) \), we conclude that with probability \( 1 - o(1) \)
\[ M \leq EM + t = 2n \exp\left( -0.072 \frac{(a - b)^2}{a + b} \right). \]

This implies that with probability \( 1 - o(1) \),
\[ |V_1' \setminus V_1| \leq 2n \exp\left( -0.072 \frac{(a - b)^2}{a + b} \right). \]

By symmetry, the same conclusion holds for \( |V_2' \setminus V_2| \).

Set 
\[ \gamma := 2 \exp\left( -0.072 \frac{(a - b)^2}{a + b} \right), \]
we have, for \( i = 1, 2 \)
\[ |V_i \cap V_i'| = n - |V_i \cap V_3^{i}'| = n - |V_3^{i'} \setminus V_3^{i}| \geq n(1 - \gamma). \]

This shows that the output \( V_1', V_2' \) form a \( \gamma \)-correct partition, with \( \gamma \) satisfying 
\[ \frac{(a - b)^2}{a + b} = \frac{1}{0.072} \log_2 \frac{2}{\gamma} \approx 13.89 \log_2 \frac{2}{\gamma}, \]
proving our claim.

Proof of Corollary 1.4. Let us sketch the proof of Corollary 1.4. 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 \( \gamma \) does not appear in the
bound). In the analysis of Correction, we require \((a-b)^2 \geq 13.89 \log \frac{2}{\gamma}\), as shown above. If \(\gamma < \epsilon\) for a sufficiently small \(\epsilon\), this assumption implies the first. Thus, Corollary holds with assumption \((a-b)^2 \geq 13.89 \log \frac{2}{\gamma}\).

The constant 13.89 comes from the fact that the partition obtained from Spectral Partition is \(\gamma\)-correct. If one improves upon \(\gamma\), one improves 13.89. In particular, there is a constant \(\delta\) such that if the first partition is \(\delta\)-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).

5 Multiple communities

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.

**Spectral Partition II.**

1. Input \(A_0\) (the \(kn \times kn\) adjacency matrix of \(G = (V, E)\)), \(a\), \(b\) and \(k\).
2. Randomly partition \(V\) into two subsets \(Y\) and \(Z\). Let \(B\) be the adjacency matrix of the bipartite graph between \(Y\) and \(Z\).
3. 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 \(Y_1, Y_2 := Y \setminus Y_1\), respectively.
4. Let \(d := a + (k-1)b\). Zero out all the rows and columns of \(A_1\) corresponding to vertices whose degree is bigger than \(c_kd\), to obtain the matrix \(A\). Here \(c_k\) is a constant dependent on \(k\).
5. Find the space spanned by \(k\) left singular vectors of \(A\), say \(W\).
6. Project the \(Ck \log k\) random columns of \(B\) indexed by \(Y_2\) on \(W\).
7. For each projected vector, identify the top \(n/2\) coordinates. These are the output clusters.

**Figure 5: Spectral Partition 2**

This algorithm actually gives a \(\gamma\)-correct approximation for the sets \(Y_2 \cap V_1, ..., Y_2 \cap V_k\). Following an argument in [Vu14], this can then be easily extended to approximate the clusters \(V_1, ..., V_k\). Details will appear in the full version of the paper.

To analyze this algorithm, we use the machinery developed so far combined with some ideas from [Vu14]. 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 \(|V_i| = n\) and for \(u \in V_i, v \in V_j\):

\[
\mathbb{P}((u, v) \in E) = \begin{cases} 
\frac{a}{n} & \text{if } i = j \\
\frac{b}{n} & \text{if } i \neq j
\end{cases}
\]
We can write, as before
\[
A = \tilde{A} + E = \tilde{A}_1 + \Delta + E,
\]
where \(\tilde{A}, \tilde{A}_1\) are the expected matrices, and \(\Delta\) is matrix containing the deleted rows and columns. Let \(\tilde{W}\) be the span of the \(k\) left singular vectors of \(\tilde{A}_1\). We can bound \(\|\Delta\| \leq 1\) by bounding the number of high degree vertices as we did before. \(E\) is given by
\[
E_{u,v} = \begin{cases} 
1 - \frac{a}{n} & \text{w.p. } \frac{a}{n} \\
\frac{a}{n} & \text{w.p. } 1 - \frac{a}{n}
\end{cases}
\]
if \(u, v \in V_i \cap Y_1\) for some \(i \in 1, ..., k\) and
\[
E_{u,v} = \begin{cases} 
1 - \frac{b}{n} & \text{w.p. } \frac{b}{n} \\
\frac{b}{n} & \text{w.p. } 1 - \frac{b}{n}
\end{cases}
\]
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 3.3 applied to \(\tilde{A}_1 - A\) gives the following result.

**Lemma 5.1** There exists a constant 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 \(\tilde{A}_1\) is \(k\), and its least non-trivial singular value \(\lambda_k(\tilde{A}_1) > C \lambda_k(A_0) = C(a - b)\). The proof of this fact is very similar to the proof of an analogous fact in section 3 of [Vu14]. This fact, combined with lemma 5.1 and an application of Davis-Kahan bound gives

**Lemma 5.2** For any \(c > 0\), there exists constants \(C_1, C_2\) such that if \(a - b > C_1 \sqrt{a + b}\) and \(a > b \geq C_2\), \(a = o(\sqrt{\frac{n}{\log n}})\), then \(\sin \angle(\tilde{W}, W) \leq c\) with probability \(1 - o(1)\).

We randomly pick \(m = \alpha k \log k\) indices from \(Y_2\) and project the corresponding columns from the matrix \(B\), say \(a_{i_1}, ..., a_{i_m}\), onto the subspace \(W\). By the coupon collector phenomenon, given \(\epsilon_1 > 0\) and setting \(\alpha\) sufficiently large, we can guarantee with probability at least \(1 - \epsilon_1\) that we pick at least one column from each of the \(k\) clusters \(V_i \cap Y_2\), for \(i = 1, ..., k\). Let \(\tilde{a}_{i_1}, ..., \tilde{a}_{i_m}\) and \(e_{i_1}, ..., e_{i_m}\) be the corresponding columns of \(A_1\) and \(E\), respectively. For a subspace \(W_0\), let \(P_{W_0}\) be the projection on to the space \(W_0\). We have,
\[
\tilde{a}_i = P_{W} \tilde{a}_i.
\]

Note that if vertex \(i \in V_{n_i} \cap Y_2\), then
\[
\tilde{a}_i(j) = \begin{cases} 
\frac{a}{n} & \text{if } j \in V_{n_i} \cap Y_1 \\
0 & \text{if } j \in V_{n_i} \cap Y_2 \\
\frac{b}{n} & \text{otherwise}
\end{cases}
\]
Therefore, if we can recover \(\tilde{a}_i\), we can identify the set \(V_{n_i} \cap Y_1\). We now argue that we can recover \(a_i\) approximately. Since \(a_i = \tilde{a}_i + e_i\), we have
\[
P_W a_i = P_W \tilde{a}_i + P_W e_i
\]
\[
= P_W \tilde{a}_i + P_W e_i + \text{err}_i
\]
\[
= \tilde{a}_i + P_W e_i + \text{err}_i,
\]
where
\[ \text{err}_i = (P_W - P_{\bar{W}}) \bar{a}_i. \]
Since \( \sin(\angle(W, W) \leq \delta_1 \), we have for any unit vector \( v \), \( \|P_W v - P_{\bar{W}} v\| \leq \delta_1 \), which in turn implies for all \( i \)
\[ \|\text{err}_i\| \leq \delta_1 \|\bar{a}_i\|. \]

Therefore, it is enough to bound \( \|P_W e_i\| \). We recall that \( k \) is a constant that does not depend on \( n \). \( W \) is \( k \) dimensional space giving \( E \|P_W e_i\|^2 \leq k\sigma^2 \). By Markov’s inequality, it follows that
\[ \mathbb{P}(\|P_W e_i\| > K\sigma k^{1/2}) \leq K^{-2}. \]
Therefore, for any \( \epsilon_2 > 0 \) by choosing \( K \) appropriately, we can say that with probability at least \( 1 - \epsilon_2 \), all randomly chosen \( m = O(k \log k) \) indices \( i_1, \ldots, i_m \) satisfy
\[ \|P_W e_{i_j}\| < K\sigma k^{1/2} \text{ for } j = 1, \ldots, m. \]
Note that \( \|\bar{a}_i\| \geq \frac{\sigma}{\sqrt{n}} \) for all \( i \) and that \( \sigma \leq \frac{\sqrt{a}}{\sqrt{n}} \). Therefore, for any \( \delta_2 > 0 \) if \( a > C \) for a big enough constant \( C \), we have that \( K\sigma k^{1/2} \leq \delta_2 \|\bar{a}_i\| \) for all \( i \).

Lemma 5.3 For any constants \( \epsilon_2 > 0 \) and \( \delta_2 > 0 \), there exists a constant \( C \) such that if \( a > C \), then
\[ \mathbb{P}(\|P_W e_{i_j}\| > \delta_2 \|\bar{a}_{i_j}\|) \leq 1 - \epsilon_2 \text{ for } j = 1, \ldots, m. \]
Recall that \( i_1, \ldots, i_m \in Y_2 \) can be any \( O(k \log k) \) vertices in \( Y_2 \). For any constants \( \delta, \epsilon > 0 \), we have with probability \( 1 - \epsilon \), \( \|P_W a_{i_j} - \bar{a}_{i_j}\| \leq \delta \|\bar{a}_{i_j}\| \) for \( j = 1, \ldots, m \). By choosing the constants appropriately, this means that for any \( c > 0 \) we can identify \( 1 - c \) fraction of the vertices in \( V_{n_{i_j}} \cap Y_2 \), for \( j = 1, \ldots, m = O(k \log k) \).

The optimal dependence of \( C_2 \) on \( \gamma \), \( C_2 = C_3 \log \frac{1}{\gamma} \) can be achieved using an extra Correction subroutine exactly as in the case \( k = 2 \). The dependence of \( C_1 \) on \( \epsilon \) follows by a careful examination of the above proof. We omit the details (which will appear in the full version of the paper).

6 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 [ABBS14], 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 \( z_e \sim \text{Bernoulli}(\epsilon) \), we have
\[ y_{i,j} = \begin{cases} 
  x_i \oplus x_j & \text{w.p. } p(1 - \epsilon) \\
  x_i \oplus x_j + 1 & \text{w.p. } pe \\
  0 & \text{w.p. } 1 - p 
\end{cases}.
\]
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
\[ \bar{y}_{i,j} := E(y_{i,j}) = p(1 - \epsilon)w_{i,j} + pe(1 - w_{i,j}) = pe + p(1 - 2\epsilon)w_{i,j}. \]
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 \( W \) corresponding to the top two eigenvalues of \( Y_0 \).
4. Compute \( v_1 \), the projection of all-ones vector on to \( W \)
5. Let \( v_2 \) be the unit vector in \( W \) perpendicular to \( v_1 \).
6. Sort the vertices according to their values in \( v_2 \), 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 6: Algorithm 3

Therefore, we can write \( y_{i,j} = \bar{y}_{i,j} + \zeta_{i,j} \), where \( \zeta_{i,j} \)'s are mean zero random variables given by

\[
\zeta_{i,j} = \begin{cases} 
-p\epsilon + (1 - p(1 - 2\epsilon))w_{i,j} & \text{w.p. } p(1 - \epsilon) \\
1 - p\epsilon - (1 + p(1 - 2\epsilon))w_{i,j} & \text{w.p. } p\epsilon \\
-p\epsilon - p(1 - 2\epsilon)w_{i,j} & \text{w.p. } 1 - p.
\end{cases}
\]

This can be rewritten in a more amenable form as follows. If \( w_{i,j} = 1 \), then

\[
\zeta_{i,j} = \begin{cases} 
1 - p(1 - \epsilon) & \text{w.p. } p(1 - \epsilon) \\
-p(1 - \epsilon) & \text{w.p. } 1 - p(1 - \epsilon)
\end{cases}
\]

and if \( w_{i,j} = 0 \), then

\[
\zeta_{i,j} = \begin{cases} 
-p\epsilon & \text{w.p. } 1 - p\epsilon \\
1 - p\epsilon & \text{w.p. } p\epsilon
\end{cases}
\]

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 \( v_1 = (1, 1, \ldots, 1) \) and \( v_2 = (1, \ldots, 1, -1, \ldots, -1) \). Let \( Y = (y_{i,j}) \) and \( E = (\zeta_{i,j}) \) be \( 2n \times 2n \) matrices. Therefore, if we can find \( v_2 \), we can identify the two blocks. This is achieved by algorithm 6 (which is essentially same as algorithm 3) which takes as input the adjacency matrix \( Y \) and the edge probability \( p \).

All we have to do now is to bound \( \|E\| \). Furthermore, we have

\[
\text{Var}(\zeta_{i,j}) \leq p.
\]

Let \( \sigma^2 := p \geq \text{Var}(\zeta_{i,j}) \) 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 3.7, so we skip the details.

**Lemma 6.1** \( 0 < \epsilon_0 \leq \epsilon < \frac{1}{2} \). Then there exist constants \( C, C_1 \) such that if \( p \geq \frac{C}{n} \), then with probability
$1 - o(1), \|Y_0 - \hat{Y}\| \leq C_1 \sigma \sqrt{n} = C_1 \sqrt{np}.$

Since the second eigenvalue of $\hat{Y}$ 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 1.6 is precisely this.

References

[ABBS14] 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.

[AK94] Noga Alon and Nabil Kahale. A spectral technique for coloring random 3-colorable graphs. In SIAM Journal on Computing, pages 346–355, 1994.

[AS04] N. Alon and J.H. Spencer. The Probabilistic Method. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2004.

[BCLS87] 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.

[Bha97] Rajendra Bhatia. Matrix Analysis, volume 169. Springer, 1997.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities. A nonasymptotic theory of independence, 2013.

[Bol01] B. Bollobas. Random Graphs. Cambridge University Press, 2001.

[CO10] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. Combinatorics, Probability and Computing, 19:227–284, 2010.

[Dav63] Chandler Davis. The rotation of eigenvectors by a perturbation. Journal of Mathematical Analysis and Applications, 6(2):159 – 173, 1963.

[DF89] 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.

[DKMZ11] 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.

[FKS89] 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.
Appendix

Proof of Lemma 3.5: This proof is essentially same as that in [FO05]. Let us first define the following sets. For \( \gamma_k := 2^k \),

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

and

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

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

\[
\sum_{H} x_i A_{i,j} y_j \leq \sum_{i,j: \gamma_i \gamma_j \geq \sqrt{d}} s_i t_j \frac{d}{n} \lambda_{i,j} \frac{\gamma_i}{\sqrt{n}} \frac{\gamma_j}{\sqrt{n}}
\]

\[
= \sqrt{d} \sum_{i,j: \gamma_i \gamma_j \geq \sqrt{d}} s_i \gamma_i t_j \gamma_j \frac{\lambda_{i,j} \sqrt{d}}{\gamma_i \gamma_j}
\]

\[
= \sqrt{d} \sum_{i,j: \gamma_i \gamma_j \geq \sqrt{d}} \alpha_i \beta_j \sigma_{i,j}.
\]
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 (7.4) as follows:

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

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

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

2. $\lambda_{i,j} \leq c_2$:

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

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

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,

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

4. We now assume that we are not in cases 1–3. Therefore, we can assume that (?? holds. We consider the following sub cases.
(a) \( \log \lambda_{i,j} > (1/4)[2 \log \gamma_j + \log(1/\beta_j)] \) implies that \( \sigma_{i,j} \alpha_i \leq 4c_3(\gamma_i/\gamma_j \sqrt{d}) \). Therefore,

\[
\sqrt{d} \sum_{i,j: \gamma_i \gamma_j \geq \sqrt{d}} \alpha_i \beta_j \sigma_{i,j} = \sqrt{d} \sum_{j} \beta_j \sum_{i: \gamma_i \gamma_j \geq \sqrt{d}} \alpha_i \sigma_{i,j} \\
\leq \sqrt{d} \sum_{j} \beta_j \sum_{i: \gamma_i \gamma_j \geq \sqrt{d}} 4c_3 \frac{\gamma_i}{\gamma_j \sqrt{d}} \\
\leq \sqrt{d} \sum_{j} \beta_j \times 8c_3 \\
\leq 32c_3 \sqrt{d}.
\]

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 \( \sigma_{i,j} \alpha_i \leq c_3 \gamma_i \sqrt{d} \). Therefore,

\[
\sqrt{d} \sum_{i,j: \gamma_i \gamma_j \geq \sqrt{d}} \alpha_i \beta_j \sigma_{i,j} = \sqrt{d} \sum_{j} \left( \beta_j \sum_{i: \gamma_i \gamma_j \geq \sqrt{d}} \alpha_i \frac{\lambda_{i,j} \sqrt{d}}{\gamma_i \gamma_j} \right) \\
\leq \sqrt{d} \sum_{j} \left( \beta_j \sum_{i: \gamma_i \gamma_j \geq \sqrt{d}} 4c_3 \frac{\gamma_i}{\sqrt{d} \gamma_j} \log \gamma_j \right) \\
\leq \sqrt{d} \left( \sum_{j} \beta_j 4c_3 \sum_{i: \gamma_i \gamma_j \geq \sqrt{d}} \frac{\gamma_i}{\sqrt{d}} \right) \\
\leq \sqrt{d} \sum_{j} \beta_j 4c_3 \times 2 \\
\leq 32c_3 \sqrt{d}.
\]

(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:

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

22
\[ \leq \sqrt{d} \sum_{i} \left( \alpha_i \sum_{j: \gamma_i \gamma_j \geq \sqrt{d}} \frac{\sqrt{d}}{\gamma_i \gamma_j} \right) \]
\[ \leq \sqrt{d} \sum_{i} (\alpha_i \times 2) \]
\[ \leq 8\sqrt{d}. \]