Sketching semidefinite programs for faster clustering

Dustin G. Mixon[*]   Kaiying Xie[†]

Abstract

Many clustering problems can be solved using semidefinite programming. Theoretical results in this vein frequently consider data with a planted clustering and a notion of signal strength such that the semidefinite program exactly recovers the planted clustering when the signal strength is sufficiently large. In practice, semidefinite programs are notoriously slow, and so speedups are welcome. In this paper, we show how to sketch a popular semidefinite relaxation of a graph clustering problem known as minimum bisection, and our analysis supports a meta-claim that the clustering task is less computationally burdensome when there is more signal.

1   Introduction

In many data science applications, one is tasked with partitioning objects into clusters so that members of a common cluster are more similar than members of different clusters. For example, given a graph, one might cluster the vertices in such a way that most edges are between vertices from a common cluster. At the same time, one ought to ensure that clusters are appropriately balanced in size, since otherwise a cluster could degenerate to a single vertex. There are a variety of graph clustering objectives that simultaneously penalize edges across clusters and varying sizes of clusters, and for each objective, there are families of graphs for which finding an optimal clustering appears to be computationally difficult. For example, given a graph with an even number of vertices, one might bisect the vertex set in such a way that minimizes the number of edges across clusters. This *minimum bisection problem* is known to be NP-hard [5].

Recently, it has been popular to demonstrate instances of the meta-claim that *clustering is only difficult when it doesn't matter*. For example, while one can encode hard instances of the traveling salesman problem as instances of minimum bisection, these sorts of instances would never appear in the context of real-world data science. Rather, a data scientist will cluster data that is meant to be clustered. With this perspective in mind, researchers have studied how a variety of clustering algorithms perform for datasets with a planted clustering structure. For example, one might attempt to solve the minimum bisection problem for random graphs drawn from the *stochastic block model*, in which edges are drawn within planted communities at a higher rate than edges across planted communities. In this setting,

---

[*]Department of Mathematics, The Ohio State University, Columbus, OH
[†]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH

it was shown in [1, 3] that a slight modification of the Goemans–Williamson semidefinite program [6] exactly recovers the planted clustering whenever it is information theoretically feasible to do so, i.e., in the regime where there are no isolated vertices with high probability; see Proposition 1 for details. Similar approaches have treated other semidefinite programming–based clustering algorithms in various settings [2, 9, 15, 13, 14].

Many of these results take the following form: "Given enough signal, the semidefinite program exactly recovers the planted clusters." In the context of graph clustering, "signal" refers to the extent to which there are more edges within clusters than across clusters (in an appropriate quantitative sense). In this paper, we pose a subtler meta-claim:

Clustering is easier when there is more signal.

Indeed, while previous results determined how much signal is necessary and sufficient for clustering to be computationally feasible, the above meta-claim suggests that the computational burden should decline gracefully with additional signal. This makes intuitive sense considering it's easier to find a needle in a haystack when the haystack contains more needles. Such behavior is particularly welcome in the context of semidefinite programming, as solvers are notoriously slow for large datasets despite having polynomial runtime. The goal of this paper is to demonstrate this meta-claim in the special case of minimum bisection by semidefinite programming under the stochastic block model. In other words, we provide a method to systematically decrease runtime for instances with more signal.

We start by formally defining the **stochastic block model**. Let $G \sim \mathsf{SBM}(n_1, n_2, p, q)$ denote a random graph with vertex set $V(G) = S_1 \sqcup S_2$ such that $|S_1| = n_1$ and $|S_2| = n_2$ and whose edges are independent Bernoulli random variables. For every pair of vertices, if they reside in the same community S_i, we draw an edge between them with probability p, and otherwise the edge probability is q. If $p > q$, then this models how a social network exhibits more connections within a community than between communities. In our problem, we do not have access to the partition $\{S_1, S_2\}$. One might randomize the communities to model this lack of information, but we do not bother with this formality here.

In the special case where $n_1 = n_2$, one may exactly recover $\{S_1, S_2\}$ from G provided p is appropriately large compared to q. To do so, we follow [1, 3] by encoding G with the matrix B defined by $B_{ij} = 1$ if $i \leftrightarrow j$, $B_{ii} = 0$, and otherwise $B_{ij} = -1$, and we then solve the program

$$\text{maximize} \quad x^\top B x \quad \text{subject to} \quad \mathbf{1}^\top x = 0, \quad x \in \{\pm 1\}^n. \tag{1}$$

This corresponds to finding the minimum bisection of G, and one may show that the optimizers of this combinatorial program take the form $\pm(1_{\hat{S}_1} - 1_{\hat{S}_2})$, where $\{\hat{S}_1, \hat{S}_2\}$ is the maximum likelihood estimator of $\{S_1, S_2\}$. In pursuit of a computationally efficient alternative, one is inclined to consider a semidefinite program obtained by lifting $X := xx^\top$ and relaxing:

$$\text{maximize} \quad \operatorname{tr}(BX) \quad \text{subject to} \quad \operatorname{diag}(X) = 1, \quad X \succeq 0. \tag{2}$$

Following [1, 3], we relaxed away the convex constraint $\mathbf{1}^\top x = 0$, but this does not hurt performance. In fact, this relaxation exactly recovers $\{S_1, S_2\}$ in the regime in which it is information theoretically feasible to do so:

Proposition 1 ([1, 3])**.** *Select $\alpha > \beta > 0$ and for each $n \in 2\mathbb{N}$, draw $G \sim \mathsf{SBM}(n/2, n/2, p, q)$ with $p = (\alpha \log n)/n$ and $q = (\beta \log n)/n$ and with planted communities $\{S_1, S_2\}$.*

(a) If $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, then (2) recovers $\{S_1, S_2\}$ from G with probability $1 - o(1)$.

(b) If $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}$, then it is impossible to recover $\{S_1, S_2\}$ from G.

This phase transition partitions the set of all (α, β) into two regimes: one in which $\{S_1, S_2\}$ can be recovered from G in polynomial time by semidefinite programming, and another in which no algorithm exists (not even an inefficient one) that recovers $\{S_1, S_2\}$ from G. Recalling our meta-claim, we would like to show that in the former case, a larger choice of α for a fixed β makes it more computationally efficient to recover $\{S_1, S_2\}$ from G. In this spirit, we first consider Figure 1. This figure illustrates that the SDP solver empirically behaves according to our meta-claim, taking less time to identify the clustering when there is more signal available. In this paper, we do not explain this specific phenomenon. Instead, we analyze a modification to the SDP algorithm using an idea that might be transferred more easily to other settings. In fact, Figure 1 also illustrates that our modification provides a substantial speedup over the original SDP.

Recall the general **sketch-and-solve approach**: Given a large problem, we

1. randomly project to a smaller version of the same problem,

2. solve the smaller version of the problem, and then

3. use the solution to the smaller problem to (approximately) solve the original problem.

Notice that step 2 promises to be faster since the size of the problem is smaller. This approach has been particularly effective in approximately solving large least squares problems [18], and some work has been done to transfer these ideas to the setting of semidefinite programs [19, 4, 20].

We will apply the sketch-and-solve approach to systematically decrease the computational burden of clustering given more signal. In the next section, we show how to perform steps 1 and 3 above, thereby reducing our task to solving step 2. Specifically, we sketch by passing to the subgraph induced by a random subset of vertices, but in doing so, we no longer have communities of equal size. This motivates the study of the unbalanced stochastic block model, and in Section 3, we show a slight modification of (2) exactly recovers the communities in this setting. In Section 4, we combine the ideas and results in Sections 2 and 3 to state and prove our main result. As we will see, we can afford to sketch down to a smaller subgraph when there is more signal, which corroborates our meta-claim. We conclude in Section 5 with a discussion.

2 Exact recovery from a sketch oracle

As discussed in the previous section, our approach is to sketch the original graph to a smaller graph, solve the clustering problem for the smaller graph, and then use these small clusters to determine a clustering for the entire graph. For the last step, we will assign each vertex to the small cluster that it shares more edges with. The following lemma indicates that this sketch-and-solve approach identifies the planted clusters provided (1) the small graph we sketch to is not too small relative to the signal (measured in terms of the sketching
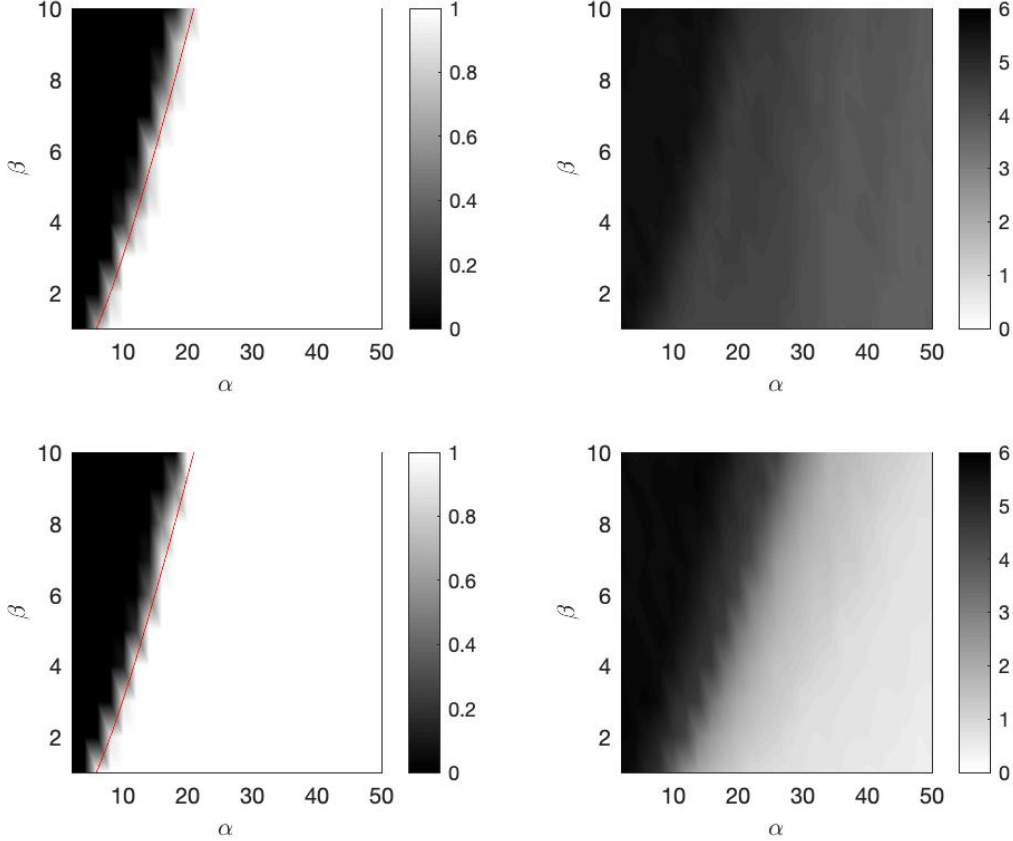
Figure 1: For each $\alpha \in \{2, 4, \ldots, 50\}$ and $\beta \in \{1, 2, \ldots, 10\}$, perform the following experiment 10 times and plot the results. Draw a random graph with distribution $\mathsf{SBM}(n_1 = 150, n_2 = 150, p = (\alpha \log n)/n, q = (\beta \log n)/n)$. Attempt to exactly recover the planted communities by solving the Abbe–Bandeira–Hall SDP (2) in CVX [7]. The proportion of recovery is displayed in **(top left)**, while the average runtime (in seconds) is displayed in **(top right)**. Next, attempt to exactly recover the planted communities using the sketch-and-solve method described in this paper; see Section 5 for details. The recovery rate and average runtime are displayed in **(bottom left)** and **(bottom right)**, respectively. The red curve depicts the phase transition from Proposition 1. For both approaches, the runtime is smaller when $\alpha \gg \beta$, and more dramatically so for the sketch-and-solve method. For example, when $\alpha = 50$ and $\beta = 1$, the average runtime for the Abbe–Bandeira–Hall SDP is 3.8170 seconds, while the sketch-and-solve method is over 8 times faster, taking an average of 0.4520 seconds.

4

parameter γ relative to the edge density parameters α and β) and (2) we correctly identify the planted small clusters R_1 and R_2. We discuss how to accomplish (2) in the next section.

Lemma 2. *Draw $G \sim \mathsf{SBM}(n/2, n/2, p, q)$ with planted communities $\{S_1, S_2\}$ and with $p = (\alpha \log n)/n$ and $q = (\beta \log n)/n$, where $\alpha > \beta > 0$. Draw vertices V at random according to a Bernoulli process with rate γ and put $R_i := S_i \cap V$ for both $i \in \{1, 2\}$. Let $e(v, S)$ denote the number of edges in G between v and $S \subseteq V(G)$ and take*

$$\hat{S}_i := R_i \cup \{v \in V(G) \setminus V : e(v, R_i) > e(v, R_{3-i})\}.$$

Then $(\hat{S}_1, \hat{S}_2) = (S_1, S_2)$ with probability $1 - o(1)$ provided

$$\gamma > \frac{8}{3} \cdot \frac{2\alpha + \beta}{(\alpha - \beta)^2}.$$

Proof. Let $\mathcal{E}$ denote the success event. After conditioning on V, the union bound gives

$$\mathbb{P}(\mathcal{E}^c | V) \leq \sum_{i \in \{1,2\}} \sum_{v \in S_i} 1_{\{v \in V(G) \setminus V\}} \cdot \mathbb{P}(\{e(v, R_i) \leq e(v, R_{3-i})\} | V).$$

Conditioned on V, then for each $v \in S_i \cap (V(G) \setminus V)$, the quantity $e(v, R_i) - e(v, R_{3-i})$ takes the form

$$\sum_{j=1}^{K_i} B_j^{(p)} - \sum_{j=1}^{K_{3-i}} B_j^{(q)},$$

where $K_i := |R_i|$ and the terms in the sums are independent Bernoulli random variables with rate indicated by the superscript. The mean of this sum is $K_i p - K_{3-i} q$. After centering, each term has absolute value at most 1 almost surely, and the variance of the sum is

$$K_i p(1 - p) + K_{3-i} q(1 - q) \leq K_i p + K_{3-i} q.$$

If in addition to $\{v \in S_i \cap (V(G) \setminus V)\}$ we restrict to the event $\mathcal{E}_1 \cap \mathcal{E}_2$, where $\mathcal{E}_i := \{K_i p - K_{3-i} q \geq 0\}$, then taking $K := \frac{K_1 + K_2}{2}$ and $J := |\frac{K_1 - K_2}{2}|$ gives $K - J = \min\{K_1, K_2\}$, $K + J = \max\{K_1, K_2\}$, and

$$\begin{aligned}
|(K - J)p - (K + J)q| &= |\min\{K_1, K_2\}p - \max\{K_1, K_2\}q| \\
&= \min\{K_1, K_2\}p - \max\{K_1, K_2\}q \\
&= \min\{K_1 p - K_2 q, K_2 p - K_1 q\} \leq K_i p - K_{3-i} q = |K_i p - K_{3-i} q|
\end{aligned}$$

for each $i \in \{1, 2\}$. Thus, on this event, we may apply Bernstein's inequality for bounded variables (see Theorem 2.8.4 in [17]):

$$\begin{aligned}
\mathbb{P}(\{e(v, R_i) - e(v, R_{3-i}) \leq 0\} | V) &\leq \exp\left(-\frac{(K_i p - K_{3-i} q)^2/2}{K_i p + K_{3-i} q + (K_i p - K_{3-i} q)/3}\right) \\
&\leq \exp\left(-\frac{(K(p - q) - J(p + q))^2/2}{K(\frac{4}{3}p + \frac{2}{3}q) + J(\frac{4}{3}p - \frac{2}{3}q)}\right).
\end{aligned}$$

5

Denote $\mathcal{E}_3 := \{J \le \epsilon K\}$ for some $\epsilon \in (0, \frac{\alpha-\beta}{\alpha+\beta})$ to be selected later. Then on the event $\{v \in S_i \cap (V(G) \setminus V)\} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, it further holds that

$$
\mathbb{P}(\{e(v, R_i) - e(v, R_{3-i}) \le 0\}|V) \le \exp\left(-\frac{(K(p-q) - J(p+q))^2/2}{K(\frac{4}{3}p + \frac{2}{3}q) + J(\frac{4}{3}p - \frac{2}{3}q)}\right)
$$
$$
\le \exp\left(-\frac{((p-q) - \epsilon(p+q))^2/2}{(\frac{4}{3}p + \frac{2}{3}q) + \epsilon(\frac{4}{3}p - \frac{2}{3}q)} \cdot K\right) =: e_{p,q,\epsilon}(K).
$$

Overall, we may bound the failure probability:

$$
\mathbb{P}(\mathcal{E}^c) = \mathbb{E}[\mathbb{P}(\mathcal{E}^c|V)] \le \mathbb{E}\left[\sum_{i \in \{1,2\}} \sum_{v \in S_i} 1_{\{v \in V(G) \setminus V\}} \cdot \mathbb{P}(\{e(v, R_i) \le e(v, R_{3-i})\}|V)\right]
$$
$$
\le \mathbb{E}\left[\sum_{i \in \{1,2\}} \sum_{v \in S_i} 1_{\{v \in V(G) \setminus V\}} \cdot \left(1_{(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)^c} + 1_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} \cdot e_{p,q,\epsilon}(K)\right)\right]
$$
$$
\le \mathbb{E}\left[\sum_{i \in \{1,2\}} \sum_{v \in S_i} \left(1_{(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)^c} + e_{p,q,\epsilon}(K)\right)\right]
$$
$$
= n\left(\mathbb{P}((\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)^c) + \mathbb{E}[e_{p,q,\epsilon}(K)]\right),
$$

where the last inequality uses $e_{p,q,\epsilon}(K) \ge 0$. We will find $\epsilon \in (0, \frac{\alpha-\beta}{\alpha+\beta})$ such that

$$
\mathbb{P}(\mathcal{E}_1^c) = e^{-\Omega(n)}, \qquad \mathbb{P}(\mathcal{E}_2^c) = e^{-\Omega(n)}, \qquad \mathbb{P}(\mathcal{E}_3^c) = e^{-\Omega(n)}, \qquad \mathbb{E}[e_{p,q,\epsilon}(K)] = o(1/n),
$$

from which it follows that $\mathbb{P}(\mathcal{E}^c) = o(1)$, as desired.

For the first three estimates, it will be helpful to first consider independent Bernoulli variables $X_1, \ldots, X_{n/2}$ and $Y_1 \ldots, Y_{n/2}$, all with rate γ. Given $a > b > 0$, we bound

$$
f(a, b) := \mathbb{P}\left\{\sum_{j=1}^{n/2} aX_j - \sum_{j=1}^{n/2} bY_j < 0\right\}.
$$

Indeed, for each $i \in \{1, 2\}$, it holds that $f(a, b)$ equals the probability that $aK_i - bK_{3-i} < 0$, since K_i is the size of the set R_i, which is drawn by a Bernoulli process of rate γ from the set S_i, which in turn is a fixed set of size $n/2$. The expectation of the above sum is $\frac{n}{2}(a-b)\gamma$, and after centering, each term in the sum has absolute value at most a almost surely. The variance of the sum is $\frac{n}{2}(a^2 + b^2)\gamma(1 - \gamma)$. As such, Bernstein's inequality for bounded variables gives

$$
f(a, b) \le \exp\left(-\frac{(\frac{n}{2}(a-b)\gamma)^2/2}{\frac{n}{2}(a^2 + b^2)\gamma(1 - \gamma) + a \cdot \frac{n}{2}(a-b)\gamma/3}\right).
$$

Notice that $\mathbb{P}(\mathcal{E}_1^c) = \mathbb{P}(\mathcal{E}_2^c) = f(p, q) = f(\frac{\alpha \log n}{n}, \frac{\beta \log n}{n})$. Simplifying then gives

$$
\mathbb{P}(\mathcal{E}_i^c) \le \exp\left(-\frac{(\frac{1}{2}(\alpha - \beta)\gamma)^2/2}{\frac{1}{2}(\alpha^2 + \beta^2)\gamma(1 - \gamma) + \alpha \cdot \frac{1}{2}(\alpha - \beta)\gamma/3} \cdot n\right),
$$

6

which is $e^{-\Omega(n)}$ since $\alpha > \beta$. Next,

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_3^c) &= \mathbb{P}\{J > \epsilon K\} \\
&= \mathbb{P}\{|K_1 - K_2| > \epsilon(K_1 + K_2)\} \\
&\leq \mathbb{P}\{K_1 - K_2 > \epsilon(K_1 + K_2)\} + \mathbb{P}\{K_2 - K_1 > \epsilon(K_1 + K_2)\} \\
&= \mathbb{P}\{(1 + \epsilon)K_2 - (1 - \epsilon)K_1 < 0\} + \mathbb{P}\{(1 + \epsilon)K_1 - (1 - \epsilon)K_2 < 0\} \\
&= 2 \cdot f(1 + \epsilon, 1 - \epsilon),
\end{aligned}
$$

which is $e^{-\Omega(n)}$ since $\epsilon \in (0, 1)$.

It remains to show that $\mathbb{E}[e_{p,q,\epsilon}(K)] = o(1/n)$. Since $K = |V|/2$, we may write

$$
e_{p,q,\epsilon}(K) = \exp\left(-\frac{((p - q) - \epsilon(p + q))^2/4}{(\frac{4}{3}p + \frac{2}{3}q) + \epsilon(\frac{4}{3}p - \frac{2}{3}q)} \cdot |V| \right).
$$

Recall that the moment generating function of a random variable X is defined by $M_X(t) := \mathbb{E}[e^{tX}]$. As such, we may interpret $\mathbb{E}[e_{p,q,\epsilon}(K)]$ as the moment generating function of $|V|$ evaluated at a point. Since $|V|$ has binomial distribution over n trials with rate γ, this gives

$$
\mathbb{E}[e_{p,q,\epsilon}(K)] = \left[1 - \gamma + \gamma \cdot \exp\left(-\frac{((p - q) - \epsilon(p + q))^2/4}{(\frac{4}{3}p + \frac{2}{3}q) + \epsilon(\frac{4}{3}p - \frac{2}{3}q)} \right) \right]^n =: ((1 - \gamma) + \gamma e^{-c(\log n)/n})^n,
$$

where $c := \frac{((\alpha - \beta) - \epsilon(\alpha + \beta))^2/4}{(\frac{4}{3}\alpha + \frac{2}{3}\beta) + \epsilon(\frac{4}{3}\alpha - \frac{2}{3}\beta)}$. Next, we compare to linear approximations to obtain

$$
\begin{aligned}
((1 - \gamma) + \gamma e^{-c(\log n)/n})^n &= \exp(n \log((1 - \gamma) + \gamma e^{-c(\log n)/n})) \\
&\leq \exp(n((1 - \gamma) + \gamma e^{-c(\log n)/n} - 1)) \\
&= \exp(-\gamma n(1 - e^{-c(\log n)/n})) = \exp\left(-\gamma n \cdot (1 - o(1)) \cdot \frac{c \log n}{n}\right).
\end{aligned}
$$

Finally, by our assumption on γ, it holds that $\gamma > 1/c$ for every sufficiently small $\epsilon > 0$. Taking any such ϵ gives that the above quantity is $o(1/n)$, as desired. $\qquad \square$

3 The unbalanced stochastic block model

Lemma 2 reduces our problem of exact recovery of balanced communities in the stochastic block model to a smaller problem. Unfortunately, by sketching to a random subgraph, the planted communities are no longer balanced, and so we cannot naively apply Proposition 1. Still, considering the overwhelming success of (2) in the balanced case, one is inclined to try some version of it in the more general case.

To this end, we first note that the choice of $B = 2A - J + I$ in (1) as an encoding of G is somewhat arbitrary; here, A denotes the adjacency matrix of G, and J denotes the all-ones matrix. Suppose we replaced B with any $\tilde{B} = aA + bJ + cI$ with $a > 0$. Then the constraints $\mathbf{1}^\top x = 0$ and $x \in \{\pm 1\}^n$ together ensure that

$$
x^\top \tilde{B} x = x^\top (aA + bJ + cI)x = a \cdot x^\top A x + c \cdot n.
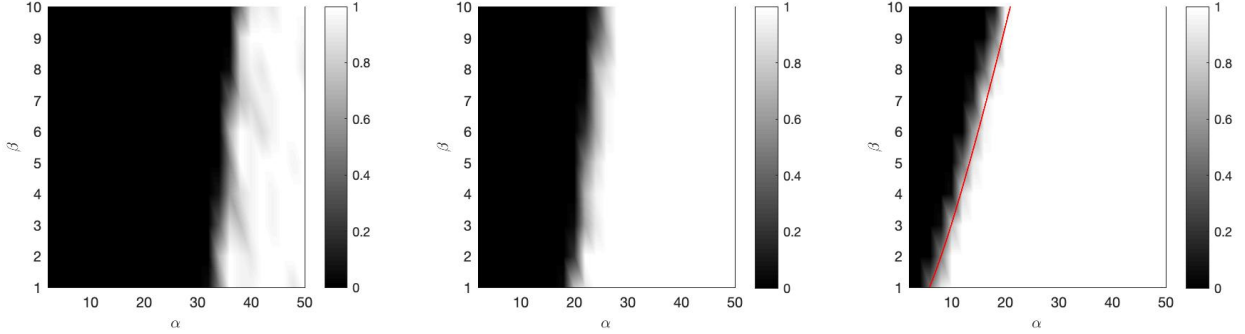$$

Figure 2: Succes rates for exact recovery under $\mathsf{SBM}(n_1 = 100, n_2 = 200, p = (\alpha \log n)/n, q = (\beta \log n)/n)$. **(left)** Goemans–Williamson SDP. **(center)** Abbe–Bandeira–Hall SDP. **(right)** Proposed SDP, which assumes access to the value of $(p + q)/2$. The red curve depicts the curve $\sqrt{\alpha} - \sqrt{\beta} = \sqrt{2}$, which by Proposition 1 is the phase transition for exact recovery in the balanced case. It appears that the same phase transition holds for the proposed SDP in the unbalanced case.

Observe that the maximizer in (1) is the same for all such objectives. However, we can expect different choices of $\tilde{B}$ to produce different optimizers once we relax to the SDP. Of course, there is no change to the SDP if we change c since X is constrained to have all-ones diagonal, and so we take $c = 0$ for simplicity. Next, we can rescale $\tilde{B}$ so that $a = 1$ without changing the SDP. Overall, we are interested in encodings of the form $\tilde{B} = A - \mu J$. The choice $\mu = \frac{1}{2}$ corresponds to the Abbe–Bandeira–Hall encoding [1]. Alternatively, we could run the Goemans–Williamson relaxation of maximum cut on the complement of G, which corresponds to taking $\mu = 1$ [6]. These two SDPs behave similarly for $G \sim \mathsf{SBM}(n_1, n_2, p, q)$ when $n_1 = n_2$, but as Figure 2 illustrates, the Abbe–Bandeira–Hall SDP performs better in the unbalanced case. Curiously, it appears that the choice $\mu = (p + q)/2$ exhibits the phase transition from Proposition 1.

In this paper, we consider the family of semidefinite programs

$$\text{maximize} \quad \operatorname{tr}((A - \mu J)X) \quad \text{subject to} \quad \operatorname{diag}(X) = 1, \quad X \succeq 0. \qquad ((A, \mu)\text{-SDP})$$

Similar to [1], we pass to the dual program to obtain an optimality condition:

Lemma 3. *Let A denote the adjacency matrix of a simple graph G on n vertices. Partition the vertices $S_1 \sqcup S_2 = V(G)$ and put $n_1 := |S_1|$, $n_2 := |S_2|$, and $g := 1_{S_1} - 1_{S_2}$. Let G^+ denote the subgraph of G with edge set $E(S_1, S_1) \cup E(S_2, S_2)$, and let G^- denote the subgraph with edge set $E(S_1, S_2)$. Finally, let D^+ and D^- denote the diagonal matrices of vertex degrees in G^+ and G^-. If the matrix*

$$D^+ - D^- - \mu(n_1 - n_2) \operatorname{diag}(g) - A + \mu J$$

is positive semidefinite with rank $n - 1$, then $gg^\top$ is the unique solution to (A, μ)-SDP.

Proof. The dual of (A, μ)-SDP is given by

$$\text{minimize} \quad \operatorname{tr}(Y) \quad \text{subject to} \quad Y \succeq A - \mu J, \quad Y \text{ diagonal}.$$

To verify weak duality, we have

$$\mathrm{tr}((A - \mu J)X) \leq \mathrm{tr}(YX) = \mathrm{tr}(Y),$$

where the first inequality follows from the semidefinite constraints, and the equality follows from the diagonal constraints. By the hypotheses of the lemma, we see that

$$X_0 := gg^\top, \qquad Y_0 := D^+ - D^- - \mu(n_1 - n_2)\,\mathrm{diag}(g)$$

are primal- and dual-feasible, respectively. As such, by rearranging the above inequality, it suffices to show that equality in

$$\mathrm{tr}((Y_0 - A + \mu J)X) \geq 0$$

holds with primal-feasible X if and only if $X = X_0$. For the "if" direction, note that

$$\mathrm{tr}(Y_0) = \mathrm{tr}(D^+) - \mathrm{tr}(D^-) - \mu(n_1 - n_2)\,\mathrm{tr}(\mathrm{diag}(g)) = g^\top A g - \mu(n_1 - n_2)^2,$$

and so

$$\mathrm{tr}((Y_0 - A + \mu J)X_0) = g^\top(Y_0 - A + \mu J)g = \mathrm{tr}(Y_0) - g^\top A g + \mu(g^\top 1)^2 = 0.$$

For the "only if" direction, put $Z := Y_0 - A + \mu J$. Then by assumption, Z is positive semidefinite, and we just showed that $g^\top Z g = 0$. It follows that $Zg = 0$. Furthermore, Z has rank $n - 1$ by assumption, and so the nullspace of Z equals the span of g. Suppose X is primal-feasible with $\mathrm{tr}(ZX) = 0$, and consider the spectral decompositions $X = \sum_i \lambda_i x_i x_i^\top$ and $Z = \sum_j \mu_j z_j z_j^\top$. Note that $0 = \mu_1 < \mu_2 \leq \mu_n$ and $z_1 = n^{-1/2} \cdot g$. Then

$$0 = \mathrm{tr}(ZX) = \sum_i \lambda_i \sum_{j>1} \mu_j |\langle x_i, z_j \rangle|^2.$$

Since each term in the outer sum is nonnegative, it must hold that for each i, either $\lambda_i = 0$ or $|\langle x_i, z_j \rangle|^2 = 0$ for every $j > 1$. Considering $\{z_j\}_{j \in [n]}$ is an orthonormal basis for $\mathbb{R}^n$, we see that $|\langle x_i, z_j \rangle|^2 = 0$ for every $j > 1$ only if x_i is a scalar multiple of z_1. On the other hand, $\{x_i\}_{i \in [n]}$ is also an orthonormal basis, and so there is at most one such x_i. It follows that X has rank at most 1, and in particular, X is a scalar multiple of $gg^\top$. The diagonal constraint on the primal-feasible X then implies that $X = gg^\top = X_0$, as desired. $\qquad \square$

Next, we apply Lemma 3 to the stochastic block model to show when the SDP exactly recovers the planted clusters. Notice the appearance of $\frac{p+q}{2}$ as a threshold on μ:

Lemma 4. *Take $n_1, n_2 \geq 1$ and $p > q > 0$, put $n := n_1 + n_2$ and $m := |n_1 - n_2|$, draw $G \sim \mathsf{SBM}(n_1, n_2, p, q)$ with planted communities $\{S_1, S_2\}$, let A denote the adjacency matrix of G, and put $g = 1_{S_1} - 1_{S_2}$. If $\mu > q$, then $gg^\top$ is the unique solution to (A, μ)-SDP with probability at least*

$$\begin{cases} 1 - 2n \exp(-\frac{3}{2} \cdot \frac{((\mu-q)n)^2}{(3p+q+2\mu)n+3(p-q)m}) & \text{if } \mu < \frac{p+q}{2} \\[2mm] 1 - 2n \exp(-\frac{3}{16} \cdot \frac{((p-q)n-(2\mu-(p+q))m)^2}{(2p+q)n+(2p-q-\mu)m}) & \text{if } \mu \geq \frac{p+q}{2}. \end{cases}$$

9

Proof. By Lemma 3, it suffices to show that with high probability, the random matrix

$$Z := D^+ - D^- - \mu(n_1 - n_2)\operatorname{diag}(g) - A + \mu J$$

is positive semidefinite with rank $n - 1$. As established in the proof of Lemma 3, it holds that $Zg = 0$ almost surely. It remains to show that the second-smallest eigenvalue $\lambda_2(Z)$ of Z is strictly positive with high probability. Weyl's inequality (see Theorem 4.3.1 in [8]) gives

$$\lambda_2(Z) = \lambda_2(\mathbb{E}Z + Z - \mathbb{E}Z) \geq \lambda_2(\mathbb{E}Z) - \|Z - \mathbb{E}Z\|_{2\to 2}.$$

To continue, we determine the exact value of $\lambda_2(\mathbb{E}Z)$, and then we use matrix Bernstein to bound $\|Z - \mathbb{E}Z\|_{2\to 2}$ in a high-probability event.

To compute $\lambda_2(\mathbb{E}Z)$, it is helpful to assume (without loss of generality) that $S_1 = \{1, \ldots, n_1\}$ and $S_2 = \{n_1 + 1, \ldots, n\}$. We first write the matrix in block form:

$$\mathbb{E}Z = \begin{bmatrix} (a_1 - b)I_{n_1} + b1_{n_1}1_{n_1}^\top & c1_{n_1}1_{n_2}^\top \\ c1_{n_2}1_{n_1}^\top & (a_2 - b)I_{n_2} + b1_{n_2}1_{n_2}^\top \end{bmatrix},$$

where the matrix entries are given by

$$a_1 = p(n_1 - 1) - qn_2 - \mu(n_1 - n_2) + \mu, \qquad b = -p + \mu,$$
$$a_2 = p(n_2 - 1) - qn_1 + \mu(n_1 - n_2) + \mu, \qquad c = -q + \mu.$$

From this block form, we see that for each $i \in \{1, 2\}$, every vector supported on S_i that is orthogonal to 1_{S_i} is an eigenvector with eigenvalue $a_i - b$. One may simplify to obtain

$$a_1 - b = (p - \mu)n_1 + (\mu - q)n_2, \qquad a_2 - b = (p - \mu)n_2 + (\mu - q)n_1.$$

Considering $g = 1_{S_1} - 1_{S_2}$ is an eigenvector with eigenvalue 0, the remaining eigenvector h must reside in the span of 1_{S_1} and 1_{S_2} and simultaneously be orthogonal to $1_{S_1} - 1_{S_2}$. As such, we may take $h = n_2 1_{S_1} + n_1 1_{S_2}$, and multiplying by $\mathbb{E}Z$ reveals that the corresponding eigenvalue is $(\mu - q)n$. Observe that this is the second-smallest eigenvalue provided it is nonnegative and $p - \mu > \mu - q$, i.e., $q \leq \mu < (p + q)/2$. On the other hand, if $\mu \geq (p + q)/2$, then the second smallest eigenvalue is the smaller of $a_1 - b$ and $a_2 - b$. Overall, we have $\lambda_2(\mathbb{E}Z) \leq 0$ if $\mu < q$ and otherwise

$$\lambda_2(\mathbb{E}Z) = \begin{cases} (\mu - q)n & \text{if } q \leq \mu < (p+q)/2, \\ (p - \mu)\max(n_1, n_2) + (\mu - q)\min(n_1, n_2) & \text{if } \mu \geq (p+q)/2. \end{cases}$$

Notice that in the case $\mu \geq (p + q)/2$, we we may simplify this expression:

$$\lambda_2(\mathbb{E}Z) = (p - \mu)\tfrac{n+m}{2} + (\mu - q)\tfrac{n-m}{2} = \tfrac{p-q}{2} \cdot n - (\mu - \tfrac{p+q}{2}) \cdot m.$$

It remains to obtain a high-probability upper bound on $\|Z - \mathbb{E}Z\|_{2\to 2}$. To accomplish this, we will apply matrix Bernstein. First, we express $Z - \mathbb{E}Z$ as a random series. To this end, for each (i, j) such that i and j belong to the same community, i.e., $(i, j) \in (S_1 \times S_1) \cup (S_2 \times S_2) = S_1^2 \cup S_2^2$, we let B_{ij}^+ denote the Bernoulli random variable with success probability p that indicates whether i and j are adjacent in the random graph G. Similarly,

for each $(i,j) \in S_1 \times S_2$, let B_{ij}^- denote the Bernoulli random variable with success probability q that indicates whether i and j are adjacent. With these definitions, it is straightforward to verify the identity

$$Z - \mathbb{E}Z = \sum_{\substack{(i,j) \in S_1^2 \cup S_2^2 \\ i < j}} (B_{ij}^+ - p)(e_i - e_j)(e_i - e_j)^\top - \sum_{\substack{i \in S_1 \\ j \in S_2}} (B_{ij}^- - q)(e_i + e_j)(e_i + e_j)^\top.$$

Each (i,j) with $i < j$ corresponds to a unique term on the right-hand side above, which we denote by M_{ij}. This yields the desired random series:

$$Z - \mathbb{E}Z = \sum_{\substack{i,j \in [n] \\ i < j}} M_{ij}.$$

Then $\mathbb{E}M_{ij} = 0$ and $\|M_{ij}\|_{2 \to 2} \leq 2$ almost surely for every $i, j \in [n]$ with $i < j$. Next,

$$\mathbb{E}M_{ij}^2 = \begin{cases} 2p(1-p)(e_i - e_j)(e_i - e_j)^\top & \text{if } (i,j) \in S_1^2 \cup S_2^2, \ i < j \\ 2q(1-q)(e_i + e_j)(e_i + e_j)^\top & \text{if } (i,j) \in S_1 \times S_2, \end{cases}$$

and so we may write the sum in block form:

$$\sum_{\substack{i,j \in [n] \\ i < j}} \mathbb{E}M_{ij}^2 = \begin{bmatrix} c_1 I_{n_1} - 2p(1-p)1_{n_1}1_{n_1}^\top & 2q(1-q)1_{n_1}1_{n_2}^\top \\ 2q(1-q)1_{n_2}1_{n_1}^\top & c_2 I_{n_2} - 2p(1-p)1_{n_2}1_{n_2}^\top \end{bmatrix},$$

where $c_1 = 2p(1-p)n_1 + 2q(1-q)n_2$ and $c_2 = 2p(1-p)n_2 + 2q(1-q)n_1$. As before, every vector supported on S_i that is orthogonal to 1_{S_i} is an eigenvector with eigenvalue c_i. The other two eigenvectors reside in the span of 1_{S_1} and 1_{S_2}. Since g is in the nullspace, the remaining eigenvector is $h = n_2 1_{S_1} + n_1 1_{S_2}$. All together, the eigenvalues are

$$\begin{cases} 2p(1-p)n_1 + 2q(1-q)n_2 & \text{with multiplicity } n_1 - 1 \\ 2p(1-p)n_2 + 2q(1-q)n_1 & \text{with multiplicity } n_2 - 1 \\ 0 & \text{with multiplicity } 1 \\ 2q(1-q)n & \text{with multiplicity } 1. \end{cases}$$

As expected, all of these eigenvalues are nonnegative. Furthermore, since $p > q$, the largest eigenvalue is

$$\left\| \sum_{\substack{i,j \in [n] \\ i < j}} \mathbb{E}M_{ij}^2 \right\|_{2 \to 2} = 2p(1-p)\max(n_1, n_2) + 2q(1-q)\min(n_1, n_2)$$

$$\leq 2p\max(n_1, n_2) + 2q\min(n_1, n_2) = (p+q)n + (p-q)m.$$

We are now ready to apply the matrix Bernstein inequality (see Theorem 1.6.2 in [16]):

$$\mathbb{P}\{\|Z - \mathbb{E}Z\|_{2 \to 2} \geq t\} \leq 2n \exp\left(-\frac{t^2/2}{(p+q)n + (p-q)m + 2t/3}\right).$$

The result then follows by taking $t := \lambda_2(\mathbb{E}Z)$ and simplifying. $\quad\square$

11

As an aside, we point out the following corollary: If we know $\frac{p+q}{2}$, then we can solve the unbalanced case by semidefinite programming, though more signal is required when the communities are less balanced (at least for this result).

Corollary 5. *Select $\alpha > \beta > 0$ and $\delta \geq 0$. For each $n_1, n_2 \in \mathbb{N}$ with $|n_1 - n_2| \leq \delta(n_1 + n_2)$, put $n = n_1 + n_2$, $p = (\alpha \log n)/n$ and $q = (\beta \log n)/n$, draw $G \sim \mathsf{SBM}(n_1, n_2, p, q)$ with planted communities $\{S_1, S_2\}$, let A denote the adjacency matrix of G, and put $g = 1_{S_1} - 1_{S_2}$. Then $gg^\top$ is the unique solution to $(A, \frac{p+q}{2})$-SDP with probability $1 - o(1)$ provided*

$$3(\alpha - \beta)^2 > 16(2\alpha + \beta) + 24(\alpha - \beta)\delta.$$

Proof. Put $m = |n_1 - n_2|$. By Lemma 4, the success probability is at least

$$1 - 2n\exp(-\tfrac{3}{16} \cdot \tfrac{((p-q)n - (2\mu - (p+q))m)^2}{(2p+q)n + (2p-q-\mu)m}) = 1 - 2n\exp(-\tfrac{3}{16} \cdot \tfrac{((p-q)n)^2}{(2p+q)n + (3/2)(p-q)m})$$

$$\geq 1 - 2n\exp(-\tfrac{3}{16} \cdot \tfrac{((p-q)n)^2}{(2p+q)n + (3/2)(p-q)\delta n})$$

$$= 1 - 2\exp(-(\tfrac{3}{16} \cdot \tfrac{(\alpha-\beta)^2}{(2\alpha+\beta) + (3/2)(\alpha-\beta)\delta} - 1) \cdot \log n),$$

which is $1 - o(1)$ by our assumption on (α, β, δ). $\qquad\square$

In the special case where $\delta = 0$, this threshold matches the guarantee provided in [1]. Perhaps surprisingly, the true phase transition appears to be independent of δ; see Figure 2.

4 Main result

We are now ready to state our main result, which combines our sketching approach with our solver for the sketched problem:

Theorem 6 (main result). *Draw $G \sim \mathsf{SBM}(n/2, n/2, p, q)$ with planted communities $\{S_1, S_2\}$ and with $p = (\alpha \log n)/n$ and $q = (\beta \log n)/n$, where $\alpha > \beta > 0$. Consider the random variable $\mu := |E|/\binom{n}{2}$, where E denotes the random edge set of G. Next, draw vertices $V \subseteq V(G)$ at random according to a Bernoulli process with rate γ, and let A denote the adjacency matrix of the random subgraph induced by V. In the event that the solution to (A, μ)-SDP is unique and takes the form $gg^\top$ for some $g \in \{\pm 1\}^V$, select $\hat{R}_1 \sqcup \hat{R}_2 = V$ such that $g = 1_{\hat{R}_1} - 1_{\hat{R}_2}$, and otherwise let $\{\hat{R}_1, \hat{R}_2\}$ be a random partition of V. Next, let $\hat{S}_i$ denote the union of $\hat{R}_i$ and the vertices in $V(G) \setminus V$ that share more edges with $\hat{R}_i$ than with $\hat{R}_{3-i}$. Provided*

$$\gamma > \frac{16}{3} \cdot \frac{2\alpha + \beta}{(\alpha - \beta)^2},$$

it holds that $\{\hat{S}_1, \hat{S}_2\} = \{S_1, S_2\}$ with probability $1 - o(1)$.

In words, if $\alpha \gg \beta$, then for small choices of γ, we can sketch down to a subgraph with approximately γn vertices before solving the SDP. Since the runtime of the SDP is sensitive to the size of the problem instance, this promises to provide a substantial speedup. To be explicit, [12, 11, 10] give that the complexity of solving an SDP to precision $\epsilon > 0$ is $O^*(m(k + m^2 + n^\omega))$, where m is the number of equality constraints, k is the total number

of nonzero entries in the constraint and cost matrices, n is the number of rows in the square matrix of decision variables, $\omega \leq 2.373$ is the matrix multiplication constant, and O^* hides factors of the form $n^{o(1)}$ and $\log^{O(1)}(n/\epsilon)$; interestingly, this complexity can be achieved with either an interior point method or a cutting plane method. For our SDP, we have $m = n$ and $k \leq n^2$, and so the complexity of solving this SDP is $O^*(n^{3.373})$. Thus, sketching the problem by a factor of $\gamma < 1$ has the effect of multiplying the runtime by $\gamma^{3.373} \ll 1$. (We note that the caption of Figure 1 describes a more modest speedup using CVX, which spends a fraction of the runtime to construct the optimization problem instance from user input before actually solving it.) For dense graphs, we can expect $\gamma \to 0$ as $n \to \infty$, but our analysis focuses on signal levels that are close to the information-theoretic limit, so we encounter constant values of γ. To prove Theorem 6, we first need a version of Lemma 4 that accounts for the randomness in our sketching process:

Lemma 7. *Draw $G \sim \mathsf{SBM}(n/2, n/2, p, q)$ with planted communities $\{S_1, S_2\}$ and with $p = (\alpha \log n)/n$ and $q = (\beta \log n)/n$, where $\alpha > \beta > 0$. Draw vertices $V \subseteq V(G)$ at random according to a Bernoulli process with rate γ. Let A denote the adjacency matrix of the random subgraph induced by V. Suppose there exists ϵ such that $\beta < \epsilon < (n/\log n)\mu < 2\alpha - \beta$, and put $\eta := (\frac{\alpha+\beta}{2} - \epsilon)_+$. Then with probability $1 - o(1)$, the solution to (A, μ)-SDP is unique and identifies $\{S_1 \cap V, S_2 \cap V\}$ provided*

$$\gamma > \frac{16}{3} \cdot \frac{2\alpha + \beta - \eta}{(\alpha - \beta - 2\eta)^2}. \tag{3}$$

Before proving this lemma, we provide the idea of the proof. There are two sources of randomness, namely, the problem instance $G \sim \mathsf{SBM}(n/2, n/2, p, q)$ and the sketching variables $B_1, \ldots, B_n \sim \mathsf{Bernoulli}(\gamma)$, all of which are independent. Consider the subgraph $G[V]$ of G induced by $V := \{i \in [n] : B_i = 1\}$. Conditioned on the event $\{V = V_0\}$, we have

$$G[V_0] \sim \mathsf{SBM}(|S_1 \cap V_0|, |S_2 \cap V_0|, p, q)$$

with planted communities $\{S_1 \cap V_0, S_2 \cap V_0\}$. By virtue of this conditioning, we may appeal to Lemma 4. Since our bound on the success probability only depends on the sizes of $S_1 \cap V$ and $S_2 \cap V$, we can simply condition on these sizes. The sizes are similar with high probability, in which case the bound in Lemma 4 gives that we succeed with high probability. Since Lemma 4 is broken up into cases, it is a somewhat technical exercise to make this last statement rigorous:

Proof of Lemma 7. Put $g := 1_{S_1 \cap V} - 1_{S_2 \cap V}$, and let $\mathcal{S}$ denote the event that $gg^\top$ is the unique solution to (A, μ)-SDP. We condition on $K_1 := |S_1 \cap V|$ and $K_2 := |S_2 \cap V|$, which are independent of each other:

$$\mathbb{P}(\mathcal{S}^c) = \sum_{k_1=0}^{n/2} \sum_{k_2=0}^{n/2} \mathbb{P}(\mathcal{S}^c | \{K_1 = k_1\} \cap \{K_2 = k_2\}) \cdot \mathbb{P}\{K_1 = k_1\} \cdot \mathbb{P}\{K_2 = k_2\}. \tag{4}$$

Notice that $K_1, K_2 \sim \mathsf{Binomial}(\frac{n}{2}, \gamma)$, and so we expect these random variables to concentrate at $\frac{\gamma n}{2}$. With this intuition, we take $\delta > 0$ (to be selected later) and denote

$$I := \{k \in \mathbb{N} : \tfrac{(\gamma-\delta)n}{2} \leq k \leq \tfrac{(\gamma+\delta)n}{2}\}.$$

We use this interval to split the sum (4) into two parts:

$$\mathbb{P}(\mathcal{S}^c) = \sum_{\substack{(k_1,k_2)\in I\times I}} + \sum_{\substack{k_1,k_2\in\{0,\dots,n/2\}\\(k_1,k_2)\notin I\times I}}$$

$$\leq \sum_{(k_1,k_2)\in I\times I} \mathbb{P}(\mathcal{S}^c|\{K_1=k_1\}\cap\{K_2=k_2\})\cdot\mathbb{P}\{K_1=k_1\}\cdot\mathbb{P}\{K_2=k_2\}$$

$$\qquad + 2\cdot\mathbb{P}\{|K-\tfrac{\gamma n}{2}|>\tfrac{\delta n}{2}\}$$

$$\leq \left(\mathbb{P}\{K\in I\}\right)^2\cdot\max_{(k_1,k_2)\in I\times I}\mathbb{P}(\mathcal{S}^c|\{K_1=k_1\}\cap\{K_2=k_2\}) + 2\cdot\mathbb{P}\{|K-\tfrac{\gamma n}{2}|>\tfrac{\delta n}{2}\}$$

$$\leq \max_{(k_1,k_2)\in I\times I}\mathbb{P}(\mathcal{S}^c|\{K_1=k_1\}\cap\{K_2=k_2\}) + 2\cdot\mathbb{P}\{|K-\tfrac{\gamma n}{2}|>\tfrac{\delta n}{2}\} \qquad (5)$$

where $K\sim\mathsf{Binomial}(\tfrac{n}{2},\gamma)$. We bound the first term above by applying Lemma 4, and we bound the second term using Bernstein's inequality.

For the first term, first assume that $\mu\geq\frac{p+q}{2}$. Then Lemma 4 gives

$$\mathbb{P}(\mathcal{S}^c|\{K_1=k_1\}\cap\{K_2=k_2\})\leq 2(k_1+k_2)\exp\left(-\tfrac{3}{16}\cdot\tfrac{((p-q)(k_1+k_2)-(2\mu-(p+q))|k_1-k_2|)^2}{(2p+q)(k_1+k_2)+(2p-q-\mu)|k_1-k_2|}\right).$$

We will select $\delta>0$ small enough so that

$$(\alpha-\beta)(\gamma-\delta)\geq(2(2\alpha-\beta)-(\alpha+\beta))\delta,$$

which in turn implies that every $k_1,k_2\in I$ satisfies

$$(p-q)(k_1+k_2)\geq(\alpha-\beta)\tfrac{\log n}{n}\cdot(\gamma-\delta)n$$
$$\geq(2(2\alpha-\beta)-(\alpha+\beta))\delta\cdot\log n\geq(2\mu-(p+q))|k_1-k_2|.$$

With this information, we may bound the exponent:

$$\tfrac{((p-q)(k_1+k_2)-(2\mu-(p+q))|k_1-k_2|)^2}{(2p+q)(k_1+k_2)+(2p-q-\mu)|k_1-k_2|}\geq\tfrac{((\alpha-\beta)(\gamma-\delta)-(2(2\alpha-\beta)-(\alpha+\beta))\delta)^2}{(2\alpha+\beta)(\gamma+\delta)+(2\alpha-\beta-\epsilon)\delta}\cdot\log n.$$

All together, for every $k_1,k_2\in I$, we have

$$\mathbb{P}(\mathcal{S}^c|\{K_1=k_1\}\cap\{K_2=k_2\})\leq 2(\gamma+\delta)\exp\left((1-\tfrac{3}{16}\cdot\tfrac{((\alpha-\beta)(\gamma-\delta)-(2(2\alpha-\beta)-(\alpha+\beta))\delta)^2}{(2\alpha+\beta)(\gamma+\delta)+(2\alpha-\beta-\epsilon)\delta})\log n\right).$$

Thanks to our assumption
$$\gamma>\tfrac{16(2\alpha+\beta-\eta)}{3(\alpha-\beta-2\eta)^2}=\tfrac{16(2\alpha+\beta)}{3(\alpha-\beta)^2},$$

this bound is $o(1)$ for every sufficiently small $\delta>0$. Next, we consider the case in which $\mu<\frac{p+q}{2}$. Put $\zeta:=\mu n/\log n$. Then Lemma 4 gives

$$\mathbb{P}(\mathcal{S}^c|\{K_1=k_1\}\cap\{K_2=k_2\})\leq 2(k_1+k_2)\exp\left(-\tfrac{3}{2}\cdot\tfrac{((\mu-q)(k_1+k_2))^2}{(3p+q+2\mu)(k_1+k_2)+3(p-q)|k_1-k_2|}\right)$$

$$\leq 2(\gamma+\delta)\exp\left((1-\tfrac{3}{2}\cdot\tfrac{((\zeta-\beta)(\gamma-\delta))^2}{(3\alpha+\beta+2\zeta)(\gamma+\delta)+3(\alpha-\beta)\delta})\log n\right)$$

$$\leq 2(\gamma+\delta)\exp\left((1-\tfrac{3}{2}\cdot\tfrac{((\epsilon-\beta)(\gamma-\delta))^2}{(3\alpha+\beta+2\epsilon)(\gamma+\delta)+3(\alpha-\beta)\delta})\log n\right),$$

14

where the last step applies $\zeta > \epsilon$ and the fact that the map $\zeta \mapsto \frac{((\zeta-\beta)(\gamma-\delta))^2}{(3\alpha+\beta+2\zeta)(\gamma+\delta)+3(\alpha-\beta)\delta}$ is increasing over $\zeta \in [\beta, \frac{\alpha+\beta}{2}]$. Thanks to the assumption

$$\gamma > \tfrac{16(2\alpha+\beta-\eta)}{3(\alpha-\beta-2\eta)^2} = \tfrac{16(2\alpha+\beta-(\frac{\alpha+\beta}{2}-\epsilon))}{3(\alpha-\beta-2(\frac{\alpha+\beta}{2}-\epsilon))^2} = \tfrac{2}{3} \cdot \tfrac{3\alpha+\beta+2\epsilon}{(\epsilon-\beta)^2},$$

this bound is $o(1)$ for every sufficiently small $\delta > 0$.

At this point, we know that the first term in (5) is $o(1)$. For the second term, note that $K - \frac{\gamma n}{2}$ is a sum of independent, mean zero random variables X_i such that $|X_i| \le 1$ almost surely and $\mathrm{Var}(X_i) = \gamma(1-\gamma) \le \gamma$. As such, Bernstein's inequality for bounded variables gives $\mathbb{P}\{|K - \frac{\gamma n}{2}| > \frac{\delta n}{2}\} \le 2\exp(-\frac{(\delta n/2)^2/2}{n\gamma^2+(\delta n/2)/3}) = o(1)$, as desired. $\qquad\square$

Next, we provide a way of estimating $\frac{p+q}{2}$ so as to select μ:

Lemma 8. *Draw $G \sim \mathsf{SBM}(n/2, n/2, p, q)$ with planted communities $\{S_1, S_2\}$ and with $p = (\alpha \log n)/n$ and $q = (\beta \log n)/n$, and consider the random variable $\mu := |E|/\binom{n}{2}$, where E denotes the random edge set of G. Then for any fixed $c > 0$, it holds that*

$$\left|\mu - \frac{p+q}{2}\right| \le \frac{c \log n}{n^{3/2}}$$

with probability $1 - o(1)$.

Proof. Let A denote the adjacency matrix of G. It is helpful to assume (without loss of generality) that $S_1 = \{1, \ldots, n_1\}$ and $S_2 = \{n_1 + 1, \ldots, n\}$. Then

$$\mathbb{E}A = \begin{bmatrix} p(J-I) & qJ \\ qJ & p(J-I) \end{bmatrix},$$

where J and I denote the all-ones and identity matrices of order $n/2$. Then

$$\mathbb{E}\mu = \tfrac{1}{n(n-1)} \sum_{i,j=1}^{n} \mathbb{E}A_{ij} = \tfrac{n-2}{2(n-1)} \cdot p + \tfrac{n}{2(n-1)} \cdot q = \tfrac{p+q}{2} - \tfrac{p-q}{2(n-1)}.$$

Also, the random variable $\frac{n(n-1)}{2}\mu$ is a sum of $\frac{n(n-1)}{2}$ independent Bernoulli random variables, $(\frac{n}{2})(\frac{n}{2} - 1)$ with mean p and $(\frac{n}{2})^2$ with mean q. After subtracting the mean, each of these random variables has absolute value at most 1 almost surely, and the variance of the sum is

$$\mathrm{Var}(\tfrac{n(n-1)}{2}\mu) = (\tfrac{n}{2})(\tfrac{n}{2} - 1)p(1-p) + (\tfrac{n}{2})^2 q(1-q) \le \tfrac{n(n-1)}{2}\max(p, q).$$

We may therefore apply Bernstein's inequality for bounded variables:

$$\mathbb{P}\{|\mu - \mathbb{E}\mu| > t\} = \mathbb{P}\{|\tfrac{n(n-1)}{2}\mu - \mathbb{E}\tfrac{n(n-1)}{2}\mu| > \tfrac{n(n-1)}{2}t\}$$

$$\le 2\exp\left(-\frac{(\tfrac{n(n-1)}{2}t)^2/2}{\tfrac{n(n-1)}{2}\max(p,q) + (\tfrac{n(n-1)}{2}t)/3}\right) = 2\exp(-\tfrac{n(n-1)}{2} \cdot \tfrac{t^2/2}{\max(p,q)+t/3}).$$

Taking $t = \frac{c\log n}{2n^{3/2}}$ then gives

$$\mathbb{P}\{|\mu - \mathbb{E}\mu| > t\} \le 2\exp(-(1 - o(1)) \cdot \tfrac{c^2}{16\max(\alpha,\beta)} \cdot \log n) = o(1).$$

All together, with probability $1 - o(1)$, it holds that

$$|\mu - \tfrac{p+q}{2}| \le |\mu - \mathbb{E}\mu| + |\mathbb{E}\mu - \tfrac{p+q}{2}| \le \tfrac{c\log n}{2n^{3/2}} + \tfrac{|\alpha-\beta|\log n}{2n(n-1)} \le \tfrac{c\log n}{n^{3/2}}. \qquad\square$$

Proof of Theorem 6. Select $\eta > 0$ such that (3) is satisfied and put $\epsilon := \frac{\alpha+\beta}{2} - \eta$. By Lemma 8, it holds that $(n/\log n)\mu \in (\epsilon, 2\alpha - \beta)$ with probability $1 - o(1)$. As such, by Lemma 7, we have $\{\hat{R}_1, \hat{R}_2\} = \{S_1 \cap V, S_2 \cap V\}$ with probability $1 - o(1)$. The result then follows from Lemma 2. $\qquad\square$

5 Discussion

In this paper, we applied the sketch-and-solve approach to systematically reduce the computational burden of solving minimum bisection by semidefinite programming when the problem instance exhibits more signal. Figure 1 illustrates that our approach provides a substantial speedup over the original semidefinite programming approach. For this figure, we selected γ assuming access to α and β using the following rule:

$$\gamma = \min\left\{1, \frac{4}{(\sqrt{\alpha} - \sqrt{\beta})^2}\right\}.$$

Based on our experiments, we expect that our sketch-and-solve approach will succeed for large n in the regime

$$\gamma > \frac{2}{(\sqrt{\alpha} - \sqrt{\beta})^2}, \tag{6}$$

which corresponds to Proposition 1 in the special case where $\gamma = 1$.

Our investigation suggests a few opportunities for future work. First, while Theorem 6 provides a sufficient condition on γ, the apparent sketching threshold (6) warrants an explanation. Also, we did not provide a method for selecting the sketching parameter γ when α and β are unknown. In practice, one can test whether a given γ is too small by running the algorithm multiple times and observing whether substantially different clusterings emerge. Presumably, one can identify a good choice of γ by performing this test for increasing values of γ, but it would be nice to have a principled approach to select γ. Next, Figure 2 indicates that the phase transition from Proposition 1 emerges when solving exact recovery under the *unbalanced* stochastic block model. A more detailed analysis is required to explain this phenomenon. To obtain this phase transition empirically, we put $\mu = \frac{p+q}{2}$, which suggests another question: Given $G \sim \mathsf{SBM}(n_1, n_2, p, q)$ with unknown n_1, n_2, p, q, how does one estimate $\frac{p+q}{2}$? In our setting, the fact that $n_1 = n_2$ before sketching helped us to estimate $\frac{p+q}{2}$ in Lemma 8. Finally, it would be interesting if sketching SDPs in other settings (such as geometric clustering) also supports our meta-claim that clustering is easier in the presence of more signal.

Acknowledgments

References

[1] E. Abbe, A. S. Bandeira, G. Hall, Exact recovery in the stochastic block model, IEEE Trans. Inform. Theory 62 (2015) 471–487.

[2] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, R. Ward, Relax, no need to round: Integrality of clustering formulations, ITCS 2015, 191–200.

[3] A. S. Bandeira, Random Laplacian matrices and convex relaxations, Found. Comput. Math. 18 (2018) 345–379.

[4] A. Bluhm, D. Stilck França, Dimensionality reduction of SDPs through sketching, Linear Algebra Appl. 563 (2019) 461–475.

[5] M. R. Garey, D. S. Johnson, Computers and intractability: A guide to the theory of NP-completeness, Vol. 174, San Francisco: Freeman, 1979.

[6] M. X. Goemans, D. P. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, JACM 42 (1995) 1115–1145.

[7] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, `http://cvxr.com/cvx`

[8] R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, 1985.

[9] T. Iguchi, D. G. Mixon, J. Peterson, S. Villar, Probably certifiably correct k-means clustering, Math. Program. 165 (2017) 605–642.

[10] H. Jiang, T. Kathuria, Y. T. Lee, S. Padmanabhan, Z. Song, A faster interior point method for semidefinite programming, FOCS 2020, 910–918.

[11] H. Jiang, Y. T. Lee, Z. Song, S. Chiu-Wai Wong, An improved cutting plane method for convex optimization, convex-concave games, and its applications, STOC 2020, 944–953.

[12] Y. T. Lee, A. Sidford, S. Chiu-Wai Wong, A faster cutting plane method and its implications for combinatorial and convex optimization, FOCS 2015, 1049–1065.

[13] X. Li, Y. Li, S. Ling, T. Strohmer, K. Wei, When do birds of a feather flock together? k-means, proximity, and conic programming, Math. Program. 179 (2020) 295–341.

[14] S. Ling, T. Strohmer, Certifying global optimality of graph cuts via semidefinite relaxation: A performance guarantee for spectral clustering, Found. Comput. Math. (2019) 1–55.

[15] D. G. Mixon, S. Villar, R. Ward, Clustering subgaussian mixtures by semidefinite programming, Inform. Inference 6 (2017) 389–415.

[16] J. A. Tropp, An Introduction to Matrix Concentration Inequalities, Found. Trends Theor. Comput. Sci. 8 (2015) 1–230.

[17] R. Vershynin, High-dimensional probability: An introduction with applications in data science, Vol. 47, Cambridge University Press, 2018.

[18] D. P. Woodruff, Sketching as a tool for numerical linear algebra, Found. Trends Theor. Comput. Sci. 10 (2014) 1–157.

[19] A. Yurtsever, M. Udell, J. A. Tropp, V. Cevher, Sketchy decisions: Convex low-rank matrix optimization with optimal storage, PMLR 54 (2017) 1188–1196.

[20] A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, V. Cevher, Scalable semidefinite programming, SIAM J. Math. Data Sci. 3 (2021) 171–200.