Linear Programming and Community Detection*

Alberto Del Pia † Aida Khajavirad ‡ Dmitriy Kunisky §

May 11, 2022

Abstract

The problem of community detection with two equal-sized communities is closely related to the minimum graph bisection problem over certain random graph models. In the stochastic block model distribution over networks with community structure, a well-known semidefinite programming (SDP) relaxation of the minimum bisection problem recovers the underlying communities whenever possible. Motivated by their superior scalability, we study the theoretical performance of linear programming (LP) relaxations of the minimum bisection problem for the same random models. We show that, unlike the SDP relaxation that undergoes a phase transition in the logarithmic average degree regime, the LP relaxation fails in recovering the planted bisection with high probability in this regime. We show that the LP relaxation instead exhibits a transition from recovery to non-recovery in the linear average degree regime. Finally, we present non-recovery conditions for graphs with average degree strictly between linear and logarithmic.

Key words: Community detection, minimum bisection problem, linear programming, metric polytope

1 Introduction

Performing *community detection* or *graph clustering* in large networks is a central problem in applied disciplines including biology, social sciences, and engineering. In community detection, we are given a network of nodes and edges, which may represent anything from social actors and their interactions, to genes and their functional cooperation, to circuit components and their physical connections. We then wish to find *communities*, or subsets of nodes that are densely connected to one another. The exact solution of such problems typically amounts to solving NP-hard graph partitioning problems; hence, practical techniques instead produce approximations based on various heuristics [39, 10, 6, 20].

The stochastic block model. Various *generative models* of random networks with community structure have been proposed as a simple testing ground for the numerous available algorithmic techniques. Analyzing the performance of algorithms in this way has the advantage of not relying on individual test cases from particular domains, and of capturing the performance on *typical* random problem instances, rather than worst-case instances where effective guarantees of performance can seldom be made. The stochastic block model (SBM) is the most widely-studied generative model for community detection. Under the SBM, nodes are assigned to one of several communities (the "planted" partition or community assignment), and two nodes are connected with a probability depending only on their communities. In the simplest case, if there is an even number n of nodes, then we may assign the nodes to two communities of size $n/2$, where nodes in the same community are connected with probability $p = p(n)$ and nodes in different communities are

*A. Del Pia is partially funded by ONR grant N00014-19-1-2322. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research. D. Kunisky is supported by ONR Award N00014-20-1-2335, a Simons Investigator Award to Daniel Spielman, and NSF grants DMS-1712730 and DMS-1719545. Part of this work was performed while D. Kunisky was with New York University.

†Department of Industrial and Systems Engineering & Wisconsin Institute for Discovery, University of Wisconsin-Madison. E-mail: delpia@wisc.edu

‡Department of Industrial and System Engineering, Lehigh University. E-mail: aida@lehigh.edu.

§Department of Computer Science, Yale University. E-mail: dmitriy.kunisky@yale.edu

connected with probability $q = q(n)$, for some $p > q$. This is the so-called *symmetric assortative* SBM with two communities.

In this paper, we study the problem of *exact recovery* (henceforth simply *recovery*) of the communities under this model. That is, we are interested in algorithms that, with high probability (i.e., probability tending to 1 as $n \to \infty$), recover the assignment of nodes to communities correctly. For this task, the estimator that is most likely to succeed is the maximum *a posteriori* (MAP) estimator, which coincides with the *minimum bisection* of the graph, the assignment with the least number of edges across communities (see Chapter 3 of [1]). Computing the minimum bisection is NP-hard [22], and the best known polynomial-time approximations have a poly-logarithmic worst-case multiplicative error in the size of the bisection [26].

Semidefinite programming relaxations. However, there is still hope to solve the recovery problem under the SBM efficiently, since that only requires an algorithm to perform well on typical random graphs. Indeed, in this setting, one recent stream of research has shown that semidefinite programming (SDP) relaxations of various community detection and graph clustering problems successfully recover communities under suitable generative models [2, 32, 25, 5, 30, 29]. Generally speaking, these works first provide deterministic sufficient conditions for a given community assignment to be the unique optimal solution of the SDP relaxation, and then show that those conditions hold with high probability under a given model. For the symmetric assortative SBM with two communities, the work [2] gave a tight characterization of the values of p and q for which *any* algorithm, regardless of computational complexity, can recover the community assignments (often called an "information-theoretic" threshold). In particular, the authors found that recovery changes from possible to impossible in the asymptotic regime $p(n), q(n) \sim \log n/n$. The authors also conjectured that an SDP relaxation in fact achieves this limit, and proved a partial result in this direction, later extended to a full proof of the conjecture by [23, 5]. These results indicate that, remarkably, the polynomial-time SDP relaxation succeeds in recovery whenever the (generally intractable) MAP estimator does, in a suitable asymptotic sense.

Linear programming relaxations. It is widely accepted that for problems of comparable size, state-of-the-art solvers for linear programming (LP) significantly outperform those for SDP in both speed and scalability. Yet, in contrast to the rich literature on SDP relaxations, LP relaxations for community detection have not been studied. In this paper, motivated by their desirable practical properties, we study the theoretical performance of LP relaxations of the minimum bisection problem for recovery under the SBM. Recently, in [16, 15, 18, 17], the authors investigate the recovery properties of LP relaxations for K-means clustering, join object matching, K-medians clustering, and Boolean tensor factorization, respectively.

Integrality gaps of LP relaxations for graph cut problems. Perhaps the most similar line of prior work concerns LP relaxations for the maximum cut and the sparsest cut problems. Poljak and Tuza [35] consider a well-known LP relaxation of maximum cut problem often referred to as the metric relaxation [19]. They show that for Erdős-Rényi graphs with edge probability $O(\mathsf{polylog}(n)/n)$, the metric relaxation yields a trivial integrality gap of 2, while for denser graphs with edge probability $\Omega(\sqrt{\log n/n})$, this LP provides smaller integrality gaps. In [4], the authors consider an LP relaxation obtained by adding a large class of inequalities to the metric relaxation. Yet, this stronger LP does not improve the trivial integrality gap of the metric relaxation for edge probability $O(\mathsf{polylog}(n)/n)$. In [14], the authors show that, for high-girth graphs, for any fixed k, the LP relaxation obtained after k rounds of the Sherali-Adams hierarchy does not improve the trivial integrality gap either. In fact, the authors of [11] prove that, for random d-regular graphs, for every $\epsilon > 0$, there exists $\gamma = \gamma(d, \epsilon) > 0$ such that integrality gap of the LP relaxation obtained after n^γ rounds of Sherali-Adams is $2 - \epsilon$. In contrast to all of these negative results, the recent works [34, 24] provide a partial "redemption" of the Sherali-Adams relaxation for approximating the maximum cut, by showing that LP relaxations obtained from $n^{O(1)}$ rounds of Sherali-Adams obtain non-trivial worst-case approximation ratios. For the sparsest cut problem, the authors of [28] give an $O(\log n)$-approximation algorithm based on the metric relaxation, and show that constant degree expander graphs in fact yield a matching integrality gap of $\Theta(\log n)$ for this relaxation.

We draw attention to the unifying feature that, in all of these results, the upper bounds produced by LPs for dense random graphs are significantly better than those for sparser random graphs. Our results describe

another setting, concerning performance on the statistical task of recovering a planted bisection rather than approximating the size of the minimum bisection, where the same phenomenon occurs.

Other algorithmic approaches. Recent results have also shown that spectral methods can achieve optimal performance in the SBM as well [1, 3]. As these methods only require estimating the leading eigenvector of a matrix, they are faster than both LP and SDP relaxations. However, it is still valuable to study the theoretical properties of LP relaxations for community detection. First, LP methods provide a general strategy for approximating a wide range of combinatorial graph problems, and understanding the behavior of LP relaxations for the minimum bisection problem sheds light on their applicability to other situations where similar spectral methods do not exist. Second, works like [33, 36] suggest that convex relaxations enjoy robustness to adversarial corruptions of the inputs for statistical problems that spectral methods do not, making it of practical interest to understand the most efficient convex relaxation algorithms for solving statistical problems like recovery in the SBM.

Our contribution. Our work serves as the first average-case analysis of recovery properties of LP relaxations for the minimum bisection problem under the SBM. Proceeding similarly to the works on SDP relaxations described above, we begin with a deterministic analysis. First, we obtain necessary and sufficient conditions for a planted bisection in a graph to be the unique minimum bisection (see Theorems 1 and 2). These conditions are given in terms of certain simple measures of within-cluster and inter-cluster connectivity. Next, we consider an LP relaxation of the minimum bisection problem obtained by outer-approximating the cut-polytope by the metric-polytope [19]. Under certain regularity assumptions on the input graph, we derive a sufficient condition under which the LP recovers the planted bisection (see Propositions 1 and 2). This condition is indeed tight in the worst-case. Moreover, we give an extension of this result to general (irregular) graphs, provided that regularity can be achieved by adding and removing edges in an appropriate way (see Theorem 4). Finally, we present a necessary condition for recovery using the LP relaxation (see Theorem 5). This condition depends on how the average distance between pairs of nodes in the graph compares to the maximum such distance, the *diameter* of the graph. These two parameters have favorable properties for our subsequent probabilistic analysis.

Next, utilizing our deterministic sufficient and necessary conditions, we perform a probabilistic analysis under the SBM. Namely, we show that if $p = p(n)$ and $q = q(n)$ are constants independent of n, then the LP recovers the planted bisection with high probability, provided that $q \leq p - \frac{1}{2}$ (see Theorem 6). Conversely, if $q > \max\{2p - 1, \frac{1}{2}(3 - p - \sqrt{(3-p)^2 - 4p})\}$, then the LP fails to recover the planted bisection with high probability (see Theorem 7.1). Thus, within the *very dense* asymptotic regime $p(n), q(n) = \Theta(1)$, there are some parameters for which the LP achieves recovery, and others for which it does not. On the other hand, in the *logarithmic* regime $p(n), q(n) = \Theta(\log n/n)$, we prove that with high probability the LP fails to recover the planted bisection (see Theorem 8). We also present a collection of non-recovery conditions for the SBM in between these two regimes; that is, when $p(n), q(n) = \Theta(n^{-\omega})$ for some $0 < \omega < 1$ (see Theorem 7.2-3). In summary, we find that LP relaxations do not have the desirable theoretical properties of their SDP counterparts under the SBM, but rather have a novel transition between recovery and non-recovery in a different asymptotic regime.

Outline. The remainder of the paper is organized as follows. In Section 2 we consider the minimum bisection problem and obtain necessary and sufficient conditions for recovery. Subsequently, we consider the LP relaxation in Section 3 and obtain necessary conditions and sufficient conditions under which the planted bisection is the unique optimal solution of this LP. In Section 4 we address the question of recovery under the SBM in various regimes. Some technical results that are omitted in the previous sections are provided in Section 5.

2 The minimum bisection problem

Let $G = (V, E)$ be a graph. A *bisection of* V is a partition of V into two subsets of equal cardinality. Clearly a bisection of V only exists if $|V|$ is even. The *cost* of the bisection is the number of edges connecting the two sets. The *minimum bisection problem* is the problem of finding a bisection of minimum cost in a given

graph. This problem is known to be NP-hard [22] and it is a prototypical graph partitioning problem which arises in numerous applications. Throughout the paper, we assume $V = \{1, \dots, n\}$. Furthermore, we denote by (i, j) an edge in E with ends $i < j$.

Let G be a given graph and assume that there is also a fixed bisection of V which is unknown to us. We refer to this fixed bisection as the *planted bisection*, and we denote it by V_1, V_2. The question that we consider in this section is the following: When does solving the minimum bisection problem recover the planted bisection? Clearly this can only happen when the planted bisection is the unique optimal solution. From now on, we say that an optimization problem *recovers the planted bisection* if its unique optimal solution corresponds to the planted bisection.

We present a necessary and sufficient condition under which the minimum bisection problem recovers the planted bisection. This condition is expressed in terms of two parameters $d_{\text{in}}, d_{\text{out}}$ of the given graph G. Roughly speaking, d_{in} is a measure of *intra-cluster connectivity* while d_{out} is a measure of *inter-cluster connectivity*. Such parameters are commonly used in order to obtain performance guarantees for various clustering-type algorithms (see for example [27, 29, 30]). These quantities will also appear naturally in the course of our construction of a dual certificate for the LP relaxation in Theorem 4. In order to formally define these two parameters, we recall that the *subgraph* of G induced by a subset U of V, denoted by $G[U]$, is the graph with node set U, and its edge set is the subset of all the edges in E with both ends in U. Furthermore, we denote by E_1 and E_2 the edge sets of $G[V_1]$ and $G[V_2]$, respectively, and we define G_0 as the graph $G_0 = (V, E_0)$, where $E_0 := E \setminus (E_1 \cup E_2)$. The parameter d_{in} is then defined as the minimum degree of the nodes of $G[V_1] \cup G[V_2]$, while the parameter d_{out} is the maximum degree of the nodes of G_0.

2.1 A sufficient condition for recovery

We first present a condition that guarantees that the minimum bisection problem recovers the planted bisection.

Theorem 1. *If $d_{in} - d_{out} > n/4 - 1$, then the minimum bisection problem recovers the planted bisection.*

Proof. We prove the contrapositive, thus we assume that the minimum bisection problem does not recover the planted bisection, and we show $d_{\text{in}} - d_{\text{out}} \leq n/4 - 1$. Let $\tilde{V}_1, \tilde{V}_2$ denote the optimal solution to the minimum bisection problem.

Let $U_1 := V_1 \cap \tilde{V}_1$. Up to switching $\tilde{V}_1$ and $\tilde{V}_2$, we can assume without loss of generality that $|U_1| \leq n/4$. Since the minimum bisection problem does not recover the planted bisection, we have that U_1 is nonempty, and we denote by t its cardinality, i.e., $t := |U_1|$. Let $U_2 := V_2 \cap \tilde{V}_2$ and note that $|U_2| = t$. We also define $W_1 := V_1 \cap \tilde{V}_2$ and $W_2 := V_2 \cap \tilde{V}_1$. In particular, we have $V_1 = U_1 \cup W_1$, $V_2 = U_2 \cup W_2$, $\tilde{V}_1 = U_1 \cup W_2$, and $\tilde{V}_2 = U_2 \cup W_1$. See Fig. 1 for an illustration.

Figure 1: Both graphs in the figure are G. The graph on the left represents the planted bisection V_1, V_2. The graph on the right represents the optimal solution of the minimum bisection problem $\tilde{V}_1, \tilde{V}_2$.

We define the following sets of edges:

$$\begin{aligned}
R_1 :=& \{e \in E : e \text{ has one endnode in } U_1 \text{ and one endnode in } W_2\},\\
R_2 :=& \{e \in E : e \text{ has one endnode in } W_1 \text{ and one endnode in } U_2\},\\
S_1 :=& \{e \in E : e \text{ has one endnode in } U_1 \text{ and one endnode in } W_1\},\\
S_2 :=& \{e \in E : e \text{ has one endnode in } U_2 \text{ and one endnode in } W_2\},\\
T_1 :=& \{e \in E : e \text{ has one endnode in } U_1 \text{ and one endnode in } U_2\},\\
T_2 :=& \{e \in E : e \text{ has one endnode in } W_1 \text{ and one endnode in } W_2\}.
\end{aligned}$$

Furthermore, let $R := R_1 \cup R_2$, $S := S_1 \cup S_2$, and $T := T_1 \cup T_2$.

The optimality of the bisection $\tilde{V}_1, \tilde{V}_2$ implies that $|S|+\cancel{|T|} \leq |R|+\cancel{|T|}$, thus $|S| \leq |R|$. Next, we obtain an upper bound on $|R|$ and a lower bound on $|S|$. By definition of d_{out}, we have $|R_1| \leq td_{\text{out}}$ and $|R_2| \leq td_{\text{out}}$, thus $|R| \leq 2td_{\text{out}}$. Consider now the set S_1. The sum of the degrees of the nodes in U_1 in the graph $G[V_1]$ is at least td_{in}, while the sum of the degrees of the nodes of the graph $G[U_1]$ is at most $t(t-1)$. Thus we have $|S_1| \geq t(d_{\text{in}} - t + 1)$. Symmetrically, $|S_2| \geq t(d_{\text{in}} - t + 1)$, thus $|S| \geq 2t(d_{\text{in}} - t + 1)$. We obtain

$$2t(d_{\text{in}} - t + 1) \leq |S| \leq |R| \leq 2td_{\text{out}} \quad \Rightarrow \quad d_{\text{in}} - t + 1 \leq d_{\text{out}}.$$

Since $t \leq n/4$, we derive $d_{\text{in}} - d_{\text{out}} \leq n/4 - 1$. □

2.2 A necessary condition for recovery

Next, we present a necessary condition for recovery of the planted bisection. To this end, we make use of the following graph theoretic lemma.

Lemma 1. *For every positive even t, and every d with $0 \leq d \leq t/2$, there exists a d-regular bipartite graph with t nodes and where each set in the bipartition contains $t/2$ nodes. Furthermore, for every positive even t, and every d with $0 \leq d \leq t-1$, there exists a d-regular graph with t nodes.*

Proof. Fix any positive even t, and let U_1, U_2 be two disjoint sets of nodes of cardinality $t/2$. We start by proving the first part of the statement. For every d with $0 \leq d \leq t/2$ we explain how to construct a d-regular bipartite graph B_d with bipartition U_1, U_2. The graph $B_{t/2}$ is the complete bipartite graph. We recursively define B_{d-1} from B_d for every $1 \leq d \leq t/2$. The graph B_d is regular bipartite of positive degree and so it has a perfect matching (see Corollary 16.2b in [37]), which we denote by M_d. The graph B_{d-1} is then obtained from B_d by removing all the edges in M_d.

Next we prove the second part of the statement. Due to the first part of the proof, we only need to construct a d-regular graph G_d with node set $U_1 \cup U_2$, for every d with $t/2+1 \leq d \leq t-1$. Let H be the graph with node set $U_1 \cup U_2$ and whose edge set consists of all the edges with both ends in the same U_i, $i = 1, 2$. Note that H is a $t/2-1$ regular graph. For every $t/2+1 \leq d \leq t-1$, the graph G_d is obtained by taking the union of the two graphs $B_{d-t/2+1}$ and H. □

We are now ready to present a necessary condition for recovery. In particular, the next theorem implies that the sufficient condition presented in Theorem 1 is tight.

Theorem 2. *Let n be a positive integer divisible by eight, and let d_{in}, d_{out} be nonnegative integers such that $d_{in} \leq n/2-1$ and $d_{out} \leq n/2$. If $d_{in} - d_{out} \leq n/4-1$, there is a graph G with n nodes and a planted bisection with parameters d_{in}, d_{out} for which the minimum bisection problem does not recover the planted bisection.*

Proof. Let $n, d_{\text{in}}, d_{\text{out}}$ satisfy the assumptions in the statement. We explain how to construct a graph G with n nodes, and the planted bisection V_1, V_2 with parameters $d_{\text{in}}, d_{\text{out}}$. Furthermore we show that the minimum bisection problem over G does not recover the planted bisection.

The notation that we use in this proof is the same used in the proof of Theorem 1 and we refer the reader to Fig. 1 for an illustration. In our instances the set V_1 is the union of the two disjoint and nonempty sets of nodes U_1, W_1. Symmetrically, the set V_2 is the union of the two disjoint sets of nodes U_2, W_2. We will always have $|U_1| = |U_2|$, thus $|W_1| = |W_2|$. In order to define our instances, we will use the sets of edges R_1, R_2, R, S_1, S_2, S, and T_1, T_2, T, as defined in the proof of Theorem 1.

By assumption, we have $d_{\text{out}} \in \{0, \ldots, n/2\}$. We subdivide the proof into two separate cases: $d_{\text{in}} \leq d_{\text{out}} - 1$ and $d_{\text{in}} \geq d_{\text{out}}$. In all cases below, the constructed instance will have $|S| \leq |R|$. Thus, for these instances, a solution of the minimum bisection problem with cost no larger than the planted bisection V_1, V_2 is given by the bisection $U_1 \cup W_2, U_2 \cup W_1$.

Case 1: $d_{in} \leq d_{out} - 1$. Let $G[V_1]$ and $G[V_2]$ be d_{in}-regular graphs. Since $|V_1|$ and $|V_2|$ are even and $0 \leq d_{\text{in}} \leq n/2 - 1$, these graphs exist due to Lemma 1. Let $T \cup R$ be the edge set of a bipartite d_{out}-regular graph on nodes $V_1 \cup V_2$ with bipartition V_1, V_2. Note that this graph exists due to Lemma 1, since $0 \leq d_{\text{out}} \leq n/2$. Then G_0 is d_{out}-regular. Let U_1 contain only one node of V_1, and U_2 contain only one node of V_2. Hence the sets W_1, W_2 contain $n/2 - 1$ nodes each. It is then simple to verify that $|S| = 2d_{\text{in}}$ and $|R| \geq 2(d_{\text{out}} - 1)$. Thus we have $|S| = 2d_{\text{in}} \leq 2(d_{\text{out}} - 1) \leq |R|$.

Case 2: $d_{in} \geq d_{out}$. In the instances that we construct in this case we have that all sets U_1, W_1, U_2, W_2 have cardinality $n/4$. We distinguish between two subcases. In the first we assume $d_{\text{in}} \leq n/4 - 1$, while in the second subcase we have $d_{\text{in}} \geq n/4$.

Case 2a: $d_{in} \geq d_{out}$ and $d_{in} \leq n/4 - 1$. Let $G[U_1], G[W_1], G[U_2], G[W_2]$ be d_{in}-regular graphs. Since $|U_1| = |W_1| = |U_2| = |W_2|$ is even and $0 \leq d_{\text{in}} \leq n/4 - 1$, these graphs exist due to Lemma 1. Let S_1, S_2 be empty. Then $G[V_1]$ and $G[V_2]$ are d_{in}-regular.

Let T_1 be the edge set of a bipartite d_{out}-regular graph on nodes $U_1 \cup U_2$ with bipartition U_1, U_2, which exists due to Lemma 1, since $0 \leq d_{\text{out}} \leq d_{\text{in}} \leq n/4 - 1$. Symmetrically, let T_2 be the edge set of a bipartite d_{out}-regular graph on nodes $W_1 \cup W_2$ with bipartition W_1, W_2. Furthermore, let R_1, R_2 be empty. Then G_0 is d_{out}-regular.

In the instance constructed we have $|S| = |R| = 0$.

Case 2b: $d_{in} \geq d_{out}$ and $d_{in} \geq n/4$. Let $G[U_1], G[W_1], G[U_2], G[W_2]$ be complete graphs, which are $(n/4 - 1)$-regular. Let S_1 be the edge set of a bipartite $(d_{\text{in}} - n/4 + 1)$-regular graph on nodes $U_1 \cup W_1$ with bipartition U_1, W_1. This bipartite graph exists due to Lemma 1 since

$$1 = \cancel{n/4} - \cancel{n/4} + 1 \leq d_{\text{in}} - n/4 + 1 \leq n/2 - 1 - n/4 + 1 = n/4,$$

where the second inequality holds by assumption of the theorem. Symmetrically, let S_2 be the edge set of a bipartite $(d_{\text{in}} - n/4 + 1)$-regular graph on nodes $U_2 \cup W_2$ with bipartition U_2, W_2. Then $G[V_1]$ and $G[V_2]$ are d_{in}-regular.

Employing Lemma 1 similarly to the previous paragraph, we let R_1 be the edge set of a bipartite $(d_{\text{in}} - n/4 + 1)$-regular graph on nodes $U_1 \cup W_2$ with bipartition U_1, W_2, and let R_2 be the edge set of a bipartite $(d_{\text{in}} - n/4 + 1)$-regular graph on nodes $U_2 \cup W_1$ with bipartition U_2, W_1. Furthermore, let T_1 be the edge set of a bipartite $(d_{\text{out}} - d_{\text{in}} + n/4 - 1)$-regular graph on nodes $U_1 \cup U_2$ with bipartition U_1, U_2. This bipartite graph exists due to Lemma 1 since

$$0 \leq d_{\text{out}} - d_{\text{in}} + n/4 - 1 \leq \cancel{d_{\text{in}}} - \cancel{d_{\text{in}}} + n/4 - 1 = n/4 - 1,$$

where the first inequality holds by assumption of the theorem. Symmetrically, let T_2 be the edge set of a bipartite $(d_{\text{out}} - d_{\text{in}} + n/4 - 1)$-regular graph on nodes $W_1 \cup W_2$ with bipartition W_1, W_2. Then G_0 is d_{out}-regular.

Note that in the instance defined we have $|S_1| = |S_2| = |R_1| = |R_2|$, thus $|S| = |R|$. □

3 Linear programming relaxation

In order to present the LP relaxation for the minimum bisection problem, we first give an equivalent formulation of this problem. Given a graph $G = (V, E)$ with n nodes, the *cut vector* corresponding to a partition of V into two sets is defined as the vector $x \in \{0,1\}^{\binom{n}{2}}$ with $x_{ij} = 0$ for any $i < j \in V$ with nodes i, j in the same set of the partition, and with $x_{ij} = 1$ for any $i < j \in V$ with nodes i, j in different sets of the partition. Throughout the paper, we use the notations x_{ij} and x_{ji} interchangeably; that is, we set $x_{ji} := x_{ij}$ for all $i < j \in V$. We define $a_{ij} = 1$ for every $(i, j) \in E$ and $a_{ij} = 0$ for every $i, j \in V$ with $(i, j) \notin E$. Then the minimum bisection problem can be written as the problem of minimizing the linear function $\sum_{(i,j)\in E} x_{ij} = \sum_{1 \leq i < j \leq n} a_{ij} x_{ij}$ over all cut vectors subject to $\sum_{1 \leq i < j \leq n} x_{ij} = n^2/4$, where the equality constraint ensures that the two sets in the partition have the same cardinality.

The most well-known LP relaxation of the minimum bisection problem is the so-called *metric relaxation*, given by

$$\begin{aligned}
\min_{x} \quad & \sum_{1\le i<j\le n} a_{ij}x_{ij} && && \text{(LP-P)}\\
\text{s.t.} \quad & x_{ij} \le x_{ik} + x_{jk} && \forall \text{ distinct } i,j,k \in [n],\ i<j && (1)\\
& x_{ij} + x_{ik} + x_{jk} \le 2 && \forall\ 1 \le i < j < k \le n && (2)\\
& \sum_{1\le i<j\le n} x_{ij} = n^2/4. && && (3)
\end{aligned}$$

Problem (LP-P) is obtained by first convexifying the feasible region of the minimum bisection problem by requiring x to be in the *cut polytope*, i.e., the convex hull of all cut vectors, while satisfying the condition that the two sets in the partition have equal size, enforced by (3). Subsequently, the cut polytope is outer-approximated by the metric polytope defined by triangle inequalities (1) and (2). From a complexity viewpoint, Problem (LP-P) can be solved in polynomial time. From a practical perspective, it is well-understood that by employing some cutting-plane method together with dual Simplex algorithm, Problem (LP-P) can be solved efficiently.

3.1 A sufficient condition for recovery

Our goal in this section is to obtain sufficient conditions under which the LP relaxation recovers the planted bisection. To this end, first, we obtain conditions under which the cut vector corresponding to the planted bisection is an optimal solution of (LP-P). Subsequently, we address the question of uniqueness. We start by constructing the dual of (LP-P); we associate variables λ_{ijk} with inequalities (1), variables μ_{ijk} with inequalities (2), and a variable ω with the equality constraint (3). It then follows that the dual of (LP-P) is given by:

$$\begin{aligned}
\max_{\lambda,\mu,\omega} \quad & -2\sum_{1\le i<j<k\le n} \mu_{ijk} - \frac{n^2}{4}\omega && && \text{(LP-D)}\\
\text{s.t.} \quad & a_{ij} + \sum_{k\in[n]\setminus\{i,j\}} (\lambda_{ijk} - \lambda_{ikj} - \lambda_{jki} + \mu_{ijk}) + \omega = 0 && 1 \le i < j \le n && (4)\\
& \lambda_{ijk} = \lambda_{jik} \ge 0, \quad \mu_{ijk} = \mu_{ikj} = \mu_{kij} \ge 0 && \forall i \ne j \ne k \in [n],\ i<j &&
\end{aligned}$$

Let $\bar{x}$ be feasible to the primal (LP-P) and $(\bar{\lambda}, \bar{\mu}, \bar{\omega})$ be feasible to the dual (LP-D). Then $\bar{x}$ and $(\bar{\lambda}, \bar{\mu}, \bar{\omega})$ are optimal, if they satisfy complementary slackness. As before let G be a graph with planted bisection V_1, V_2. Let $\bar{x}$ be the cut vector corresponding to the planted bisection. Without loss of generality, we assume $V_1 = \{1, \ldots, \frac{n}{2}\}$, $V_2 = \{\frac{n}{2}+1, \ldots, n\}$. Then we have $\bar{x}_{ij} = 0$ for all $i, j \in V_1$ and all $i, j \in V_2$ with $i < j$, and $\bar{x}_{ij} = 1$ for all $i \in V_1$ and $j \in V_2$. For notational simplicity, in the remainder of this paper we let $G_1 := G[V_1]$ and $G_2 := G[V_2]$.

3.1.1 Regular graphs

In the following, we consider the setting where G_1 and G_2 are both d_{in}-regular graphs for some $d_{\text{in}} \in \{1, \ldots, \frac{n}{2} - 1\}$ and that G_0 is d_{out}-regular for some $d_{\text{out}} \in \{0, \ldots, \frac{n}{2}\}$, where we assume $n \ge 4$. This restrictive assumption significantly simplifies the optimality conditions and enables us to obtain the dual certificate in closed-form. In the next section, we relax this regularity assumption.

Proposition 1. *Let G_1 and G_2 be d_{in}-regular for some $d_{in} \in \{1, \ldots, \frac{n}{2} - 1\}$ and let G_0 be d_{out}-regular for some $d_{out} \in \{0, \ldots, \frac{n}{2}\}$. Then the cut vector $\bar{x}$ corresponding to the planted bisection is an optimal solution of* (LP-P) *if*

$$d_{in} - d_{out} \ge \frac{n}{4} - 1. \tag{5}$$

Proof. We prove the statement by constructing a dual feasible point $(\bar{\lambda}, \bar{\mu}, \bar{\omega})$ that together with $\bar{x}$ satisfies complementary slackness. By complementary slackness, at any such point we have $\bar{\lambda}_{ijk} = 0$ for all $i, j \in V_1$,

$k \in V_2$ and for all $i, j \in V_2$, $k \in V_1$ and $\bar{\mu}_{ijk} = 0$ for all $i, j, k \in V_1$ and all $i, j, k \in V_2$. Additionally, we make the following simplifying assumptions:

- $\bar{\lambda}_{ijk} = 0$ for all $i, j, k \in V_1$ and for all $i, j, k \in V_2$,
- $\bar{\lambda}_{ikj} = \bar{\lambda}_{jki} = 0$ for all $i, j \in V_1$ such that $(i, j) \notin E_1$ and for all $k \in V_2$, and $\bar{\lambda}_{kij} = \bar{\lambda}_{kji} = 0$ for all $i, j \in V_2$ such that $(i, j) \notin E_2$ and for all $k \in V_1$,
- $\bar{\mu}_{ijk} = 0$ for all $(i, j) \in E_1$ and $k \in V_2$, and $\bar{\mu}_{ijk} = 0$ for all $(i, j) \in E_2$ and $k \in V_1$.

It then follows that the linear system (4) simplifies to following set of equalities:

(I) For any $i < j \in V_1$ and $(i, j) \notin E_1$:
$$\sum_{k \in V_2} \bar{\mu}_{ijk} + \bar{\omega} = 0.$$

(II) For any $i < j \in V_2$ and $(i, j) \notin E_2$:
$$\sum_{k \in V_1} \bar{\mu}_{ijk} + \bar{\omega} = 0.$$

(III) For any $i \in V_1$ and $j \in V_2$ such that $(i, j) \notin E_0$:
$$\sum_{\substack{k \in V_1: \\ (i,k) \in E_1}} (\bar{\lambda}_{ijk} - \bar{\lambda}_{kji}) + \sum_{\substack{k \in V_2: \\ (j,k) \in E_2}} (\bar{\lambda}_{ijk} - \bar{\lambda}_{ikj}) + \sum_{\substack{k \in V_1: \\ (i,k) \notin E_1}} \bar{\mu}_{ijk} + \sum_{\substack{k \in V_2: \\ (j,k) \notin E_2}} \bar{\mu}_{ijk} + \bar{\omega} = 0.$$

(IV) For any $(i, j) \in E_0$:
$$1 + \sum_{\substack{k \in V_1: \\ (i,k) \in E_1}} (\bar{\lambda}_{ijk} - \bar{\lambda}_{kji}) + \sum_{\substack{k \in V_2: \\ (j,k) \in E_2}} (\bar{\lambda}_{ijk} - \bar{\lambda}_{ikj}) + \sum_{\substack{k \in V_1: \\ (i,k) \notin E_1}} \bar{\mu}_{ijk} + \sum_{\substack{k \in V_2: \\ (j,k) \notin E_2}} \bar{\mu}_{ijk} + \bar{\omega} = 0.$$

(V) For any $(i, j) \in E_1$:
$$1 - \sum_{k \in V_2} (\bar{\lambda}_{ikj} + \bar{\lambda}_{jki}) + \bar{\omega} = 0.$$

(VI) For any $(i, j) \in E_2$:
$$1 - \sum_{k \in V_1} (\bar{\lambda}_{kij} + \bar{\lambda}_{kji}) + \bar{\omega} = 0.$$

First, to satisfy condition (I) (resp. condition (II)), for each $i < j \in V_1$ such that $(i, j) \notin E_1$ and $k \in V_2$ (resp. $i < j \in V_2$ such that $(i, j) \notin E_2$ and $k \in V_1$), let
$$\bar{\mu}_{ijk} = -\big(\frac{a_{ik} + a_{jk}}{2}\big) \frac{\bar{\omega}}{d_{\text{out}}}. \tag{6}$$

Substituting (6) in condition (I) yields
$$-\frac{\bar{\omega}}{2d_{\text{out}}} \sum_{k \in V_2} (a_{ik} + a_{jk}) + \bar{\omega} = -\frac{\bar{\omega}}{2d_{\text{out}}} 2d_{\text{out}} + \bar{\omega} = 0,$$
where the first equality follows from d_{out}-regularity of G_0. It then follows that to satisfy $\bar{\mu}_{ijk} \geq 0$, we must have
$$\bar{\omega} \leq 0. \tag{7}$$
We will return to this condition once we determine $\bar{\omega}$. Next, to satisfy conditions (III), for each $(i, j) \in E_1$ and $k \in V_2$, let
$$\bar{\lambda}_{ikj} - \bar{\lambda}_{jki} = \big(\frac{a_{ik} - a_{jk}}{2}\big) \frac{\bar{\omega}}{d_{\text{out}}}, \tag{8}$$

and for each $(i,j) \in E_2$ and $k \in V_1$, let

$$\bar{\lambda}_{kij} - \bar{\lambda}_{kji} = \big(\frac{a_{ik} - a_{jk}}{2}\big)\frac{\bar{\omega}}{d_{\text{out}}}. \tag{9}$$

Substituting (6), (8), and (9) in condition (III), for each $i \in V_1$ and $j \in V_2$ such that $(i,j) \notin E_0$, we obtain

$$-\frac{\bar{\omega}}{2d_{\text{out}}}\sum_{k\in V_1} a_{jk} - \frac{\bar{\omega}}{2d_{\text{out}}}\sum_{k\in V_2} a_{ik} + \bar{\omega} = -\frac{\bar{\omega}}{2} - \frac{\bar{\omega}}{2} + \bar{\omega} = 0,$$

where the first equality follows from d_{out}-regularity of G_0.

Next we choose $\bar{\omega}$ so that condition (IV) is satisfied. Substituting (6),(8), and (9) in condition (IV), for each $(i,j) \in E_0$ we obtain

$$\begin{aligned}
&1 + \frac{\bar{\omega}}{2d_{\text{out}}}\Big(\sum_{\substack{k\in V_1,\\(i,k)\in E_1}}(1-a_{jk}) + \sum_{\substack{k\in V_2,\\(j,k)\in E_2}}(1-a_{ik}) - \sum_{\substack{k\in V_1,\\(i,k)\notin E_1}}(1+a_{jk}) - \sum_{\substack{k\in V_2,\\(j,k)\notin E_2}}(1+a_{ik})\Big) + \bar{\omega}\\
=&1 + \frac{\bar{\omega}}{2d_{\text{out}}}\Big(\sum_{\substack{k\in V_1,\\(i,k)\in E_1}} 1 - \sum_{\substack{k\in V_1,\\(i,k)\notin E_1}} 1 - \sum_{k\in V_1\setminus\{i\}} a_{jk} + \sum_{\substack{k\in V_2,\\(j,k)\in E_2}} 1 - \sum_{\substack{k\in V_2,\\(j,k)\notin E_2}} 1 - \sum_{k\in V_2\setminus\{j\}} a_{ik}\Big) + \bar{\omega}\\
=&1 + \frac{\bar{\omega}}{2d_{\text{out}}}\Big(d_{\text{in}} - (\frac{n}{2} - 1 - d_{\text{in}}) - d_{\text{out}} + 1 + d_{\text{in}} - (\frac{n}{2} - 1 - d_{\text{in}}) - d_{\text{out}} + 1\Big) + \bar{\omega}\\
=&1 + \frac{\bar{\omega}}{d_{\text{out}}}(2d_{\text{in}} + 2 - d_{\text{out}} - \frac{n}{2}) + \bar{\omega} = 0,
\end{aligned}$$

where the second equality follows from d_{in}-regularity of G_1, G_2 and d_{out}-regularity of G_0. Hence, to satisfy condition (IV) we must have:

$$\bar{\omega} = \frac{2d_{\text{out}}}{n - 4d_{\text{in}} - 4}. \tag{10}$$

It then follows that to satisfy condition (7) we must have $d_{\text{in}} \geq \frac{n}{4} - 1$, which is clearly implied by (5).

To complete the proof, by symmetry, it suffices to find nonnegative $\bar{\lambda}$ satisfying condition (V) together with relation (8). For each $(i,j) \in E_1$ and for each $k \in V_2$, letting

$$\bar{\lambda}_{ikj} = \max\Big\{0,\ \big(\frac{a_{ik} - a_{jk}}{2}\big)\frac{\bar{\omega}}{d_{\text{out}}}\Big\} + \gamma_{ij}, \quad \bar{\lambda}_{jki} = \max\Big\{0,\ \big(\frac{a_{jk} - a_{ik}}{2}\big)\frac{\bar{\omega}}{d_{\text{out}}}\Big\} + \gamma_{ij}, \tag{11}$$

for some $\gamma_{ij} \geq 0$, it follows that condition (V) can be satisfied if

$$-\frac{\bar{\omega}}{2d_{\text{out}}}\sum_{k\in V_2} |a_{ik} - a_{jk}| \leq 1 + \bar{\omega}.$$

By d_{out}-regularity of G_0, we have $\sum_{k\in V_2} |a_{ik} - a_{jk}| \leq 2d_{\text{out}}$ for all $i, j \in V_1$; hence the above inequality holds if

$$1 + 2\bar{\omega} \geq 0. \tag{12}$$

By (10), inequality (12) is equivalent to condition (5) and this completes the proof. □

We now provide a sufficient condition under which the cut vector corresponding to the planted bisection is the *unique* optimal solution of (LP-P). In [31], Mangasarian studies necessary and sufficient conditions for a solution of an LP to be unique. Essentially, he shows that an LP solution is unique if and only if it remains a solution to each LP obtained by an arbitrary but sufficiently small perturbation of its objective function. Subsequently, he gives a number of equivalent characterizations, one of which, stated below, turned out to be useful in our context.

Theorem 3 (Theorem 2, part (iv) in [31])**.** *Consider the linear program:*

$$\begin{aligned}
\min_x \quad & p^T x\\
\text{s.t.} \quad & Ax = b, \quad Cx \geq d,
\end{aligned}$$

where p, b and d are vectors in $\mathbb{R}^n$, $\mathbb{R}^m$ and $\mathbb{R}^k$ respectively, and A and C are $m \times n$ and $k \times n$ matrices respectively. The dual of this linear program is given by:

$$\begin{aligned} \max_{u,v} \quad & b^T u + d^T v \\ \text{s. t.} \quad & A^T u + C^T v = p, \quad v \geq 0. \end{aligned}$$

Let $\bar{x}$ be an optimal solution of the primal and let $(\bar{u}, \bar{v})$ be an optimal solution of the dual. Let C_i denote the ith row of C. Define $K = \{i : C_i \bar{x} = d_i, \bar{v}_i > 0\}$ and $L = \{i : C_i \bar{x} = d_i, \bar{v}_i = 0\}$. Let C_K and C_L, be matrices whose rows are C_i, $i \in K$ and $i \in L$, respectively. Then $\bar{x}$ is a unique optimal solution if and only if there exist no x different from the zero vector satisfying

$$Ax = 0, \quad C_K x = 0, \quad C_L x \geq 0.$$

Using the above characterization, we now present a sufficient condition for uniqueness of the optimal solution of (LP-P).

Proposition 2. *Let G_1 and G_2 be d_{in}-regular and let G_0 be d_{out}-regular. Then the LP relaxation recovers the planted bisection, if*

$$d_{in} - d_{out} \geq \frac{n}{4}. \tag{13}$$

Proof. We start by characterizing the index sets K and L defined in the statement of Theorem 3 for Problem (LP-P). By (6), $\bar{\mu}_{ijk} > 0$ for the following set of triplets (i, j, k):

1. all $i, j \in V_1$ such that $(i, j) \notin E_1$ and for all $k \in V_2$ such that $(i, k) \in E_0$ or $(j, k) \in E_0$
2. all $i, j \in V_2$ such that $(i, j) \notin E_2$ and for all $k \in V_1$ such that $(i, k) \in E_0$ or $(j, k) \in E_0$,

and $\bar{\mu}_{ijk} = 0$, otherwise. Moreover, by condition (13), we have $\gamma_{ij} > 0$ which in turn implies $\bar{\lambda}_{ikj}, \bar{\lambda}_{jki} > 0$ for all $(i, j) \in E_1$, $k \in V_2$ and by symmetry $\bar{\lambda}_{kij}, \bar{\lambda}_{kji} > 0$ for all $(i, j) \in E_2$, $k \in V_1$ and $\bar{\lambda}$s equal zero, otherwise.

Hence, by Theorem 3 and the proof of Proposition 1 it suffices to show that there exists no x different from the zero vector satisfying the following

(i) $\sum_{1 \leq i < j \leq n} x_{ij} = 0$.

(ii) For each $(i, j) \in E_1$ and each $k \in V_2$ (symmetrically for each $(i, j) \in E_2$ and each $k \in V_1$): $x_{ik} = x_{ij} + x_{jk}$, $x_{jk} = x_{ij} + x_{ik}$ and $x_{ij} + x_{jk} + x_{ik} \leq 0$.

(iii) For each $(i, j) \notin E_1$ and each $k \in V_2$ such that $(i, k) \in E_0$ or $(j, k) \in E_0$ (symmetrically, for each $(i, j) \notin E_2$ and each $k \in V_1$ such that $(i, k) \in E_0$ or $(j, k) \in E_0$): $x_{ij} + x_{jk} + x_{ik} = 0$, $x_{ik} \leq x_{ij} + x_{jk}$, $x_{jk} \leq x_{ij} + x_{ik}$.

To obtain a contradiction, assume that there exists a nonzero x satisfying conditions (i)-(iii). From condition (ii) it follows that

$$x_{ij} = 0, \quad \forall (i, j) \in E_1 \cup E_2, \tag{14}$$

and for each $(i, j) \in E_1$ (resp. $(i, j) \in E_2$) we have $x_{ik} = x_{jk} \leq 0$ for all $k \in V_2$ (resp. $k \in V_1$). Recall that a Hamiltonian cycle is a cycle that visits each node exactly once and a graph is called Hamiltonian if it has a Hamiltonian cycle. Dirac (1952) proved that a simple graph with m nodes ($m \geq 3$) is Hamiltonian if every node has degree $\frac{m}{2}$ or greater. By assumption (13), we have $d_{\text{in}} \geq \frac{n}{4}$. Since G_1 and G_2 are d_{in}-regular graphs with $\frac{n}{2}$ nodes, by Dirac's result, they are both Hamiltonian. Now consider a Hamiltonian cycle in G_1 (resp. G_2) denoted by $v_1 v_2 \dots v_{n/2} v_1$. It then follows that for each $k \in V_2$ we have $x_{v_i k} = x_{v_j k}$ for all $i, j \in \{1, \dots, n/2\}$. Symmetrically, for each $k \in V_1$, $x_{u_i k} = x_{u_j k}$ for all $i, j \in \{1, \dots, n/2\}$, where $u_1 u_2 \dots u_{n/2} u_1$ denotes a Hamiltonian cycle in G_2. Consequently, we have

$$x_{ij} = \alpha \leq 0, \quad \forall i \in V_1, \ j \in V_2. \tag{15}$$

From condition (iii), for each $(i,j) \notin E_1$ and each $k \in V_2$ such that $(i,k) \in E_0$ or $(j,k) \in E_0$, we have $x_{ij} = -x_{ik} - x_{jk} = -2\alpha$, where the second equality follows from (15). By d_{out}-regularity of G_0 and also by symmetry, we conclude that

$$x_{ij} = -2\alpha, \quad \forall (i,j) : i < j \in V_1,\ (i,j) \notin E_1,\ \text{or } i < j \in V_2,\ (i,j) \notin E_2 \tag{16}$$

Finally substituting (14), (15), and (16) into condition (i), we obtain

$$\left(\frac{n}{2}\right)^2 \alpha - \left(\frac{n}{2}\right)\left(\frac{n}{2} - d_{\text{in}} - 1\right)(2\alpha) = \alpha n\left(d_{\text{in}} + 1 - \frac{n}{4}\right) = 0,$$

which by assumption (13) holds only if $\alpha = 0$. However this implies that x has to be the zero vector, which contradicts our assumption. Hence, under assumption (13), the LP recovers the planted bisection. □

The results of Proposition 2 and Theorem 2 indicate that the worst-case recovery condition for the LP relaxation is *tight*, provided that the input graph is characterized in terms of $n, d_{\text{in}}, d_{\text{out}}$, G_1 and G_2 are d_{in}-regular, and G_0 is d_{out}-regular. Indeed, this is a surprising result as the triangle inequalities constitute a very small fraction of the facet-defining inequalities for the cut polytope.

In [7] the authors present a sufficient condition under which the minimum ratio-cut problem recovers a planted partition, in terms of the spectrum of the adjacency matrix of G. It can be shown that a similar recovery condition can also be obtained for the minimum bisection problem. Such a condition yields stronger recovery guarantees than that of Theorem 1 for sparse random graphs. At the time of this writing, we are not able to construct a dual certificate using the spectrum of the adjacency matrix, and this remains an interesting subject for future research.

3.1.2 General graphs

We now relax the regularity assumptions of the previous section. To this end, we make use of the next two lemmata. In the following we characterize a graph G by the three subgraphs (G_0, G_1, G_2), as defined before.

Lemma 2. *Suppose that the LP relaxation recovers the planted bisection for (G_0, G_1, G_2). Then it also recovers the planted bisection for $(\tilde{G}_0, G_1, G_2)$ where $E(\tilde{G}_0) \subset E(G_0)$.*

Proof. Let $|E(G_0) \setminus E(\tilde{G}_0)| = m$ for some $m \geq 1$ and denote by $\bar{x}$ the cut vector corresponding to the planted bisection of (G_0, G_1, G_2). Since the LP relaxation recovers the planted bisection for (G_0, G_1, G_2), the optimal value of this LP is equal to $|E(G_0)|$. Moreover, the objective value of the LP relaxation for $(\tilde{G}_0, G_1, G_2)$ at $\bar{x}$ is equal to $|E(G_0)| - m$. If $\bar{x}$ is not uniquely optimal for the latter LP, there exists a different solution $\tilde{x}$ that gives an objective value of $|E(G_0)| - m - \delta$ for some $\delta \geq 0$. Now let us compute the objective value of the LP relaxation for (G_0, G_1, G_2) at $\tilde{x}$; we get $\sum_{e \in E(G_1)} \tilde{x}_e + \sum_{e \in E(G_2)} \tilde{x}_e + \sum_{e \in E(\tilde{G}_0)} \tilde{x}_e + \sum_{e \in E(G_0) \setminus E(\tilde{G}_0)} \tilde{x}_e = |E(G_0)| - m - \delta + \sum_{e \in E(G_0) \setminus E(\tilde{G}_0)} \tilde{x}_e$. Since the planted bisection is recovered for (G_0, G_1, G_2) we must have $|E(G_0)| - m - \delta + \sum_{e \in E(G_0) \setminus E(\tilde{G}_0)} \tilde{x}_e > |E(G_0)|$; i.e., $\sum_{e \in E(G_0) \setminus E(\tilde{G}_0)} \tilde{x}_e > m + \delta$. However this is not possible since we have $x_{ij} \leq 1$ for all i, j. □

Lemma 3. *Suppose that the LP relaxation recovers the planted bisection for (G_0, G_1, G_2). Then it also recovers the planted bisection for $(G_0, \tilde{G}_1, \tilde{G}_2)$ where $E(\tilde{G}_1) \supseteq E(G_1)$ and $E(\tilde{G}_2) \supseteq E(G_2)$.*

Proof. Denote by $\bar{x}$ the cut vector corresponding to the planted bisection of (G_0, G_1, G_2). Since the LP relaxation recovers the planted bisection for (G_0, G_1, G_2), the optimal value of the LP is equal to $|E(G_0)|$. It then follows that $\bar{x}$ gives an objective value of $|E(G_0)|$ for the LP corresponding to $(G_0, \tilde{G}_1, \tilde{G}_2)$ as well. If $\bar{x}$ is not uniquely optimal for the latter LP, it means there exists a different solution $\tilde{x}$ that gives an objective value of $|E(G_0)| - \delta$ for some $\delta \geq 0$. Now compute the objective value of the LP for (G_0, G_1, G_2) at $\tilde{x}$; we obtain $\sum_{e \in E(G_1)} \tilde{x}_e + \sum_{e \in E(G_2)} \tilde{x}_e + \sum_{e \in E(G_0)} \tilde{x}_e = \sum_{e \in E(\tilde{G}_1)} \tilde{x}_e - \sum_{e \in E(\tilde{G}_1) \setminus E(G_1)} \tilde{x}_e + \sum_{e \in E(\tilde{G}_2)} \tilde{x}_e - \sum_{e \in E(\tilde{G}_2) \setminus E(G_2)} \tilde{x}_e + \sum_{e \in E(G_0)} \tilde{x}_e = |E(G_0)| - \delta - \sum_{e \in E(\tilde{G}_1) \setminus E(G_1)} \tilde{x}_e - \sum_{e \in E(\tilde{G}_2) \setminus E(G_2)} \tilde{x}_e$. Since the planted bisection is recovered for (G_0, G_1, G_2) we must have $|E(G_0)| - \delta - \sum_{e \in E(\tilde{G}_1) \setminus E(G_1)} \tilde{x}_e - \sum_{e \in E(\tilde{G}_2) \setminus E(G_2)} \tilde{x}_e > |E(G_0)|$; i.e., $\sum_{e \in E(\tilde{G}_1) \setminus E(G_1)} \tilde{x}_e + \sum_{e \in E(\tilde{G}_2) \setminus E(G_2)} \tilde{x}_e < -\delta$. However this is not possible since we have $x_{ij} \geq 0$ for all i, j. □

Utilizing Lemma 2, Lemma 3 and Proposition 2, we now present a sufficient condition for recovery of the planted bisection in general graphs. As in Section 2, we denote by d_{in} the minimum node degree of $G_1 \cup G_2$ and we denote by d_{out} the maximum node degree of G_0.

Theorem 4. *Assume that G_1 and G_2 contain d_{in}-regular subgraphs on the same node set and assume that G_0 is a subgraph of a d_{out}-regular bipartite graph with the same bipartition. Then the LP relaxation recovers the planted bisection, if $d_{in} - d_{out} \geq \frac{n}{4}$.*

We should remark that in general, the assumptions of Theorem 4 are restrictive; consider a graph G_1 with the minimum node degree d_{in}. Then it is simple to construct instances which do not contain a d_{in}-regular subgraph on the same node set. Similarly, in general, a bipartite graph G_0 with maximum node degree d_{out} is not a subgraph of a d_{out}-regular bipartite graph with the same bipartition. However, as we show in Section 4, for certain random graphs, these assumptions hold with high probability.

3.2 A necessary condition for recovery

Consider Problem (LP-P); in this section, we present a necessary condition for the recovery of the planted bisection using this LP relaxation. To this end, we construct a feasible point of the LP whose corresponding objective value, under certain conditions, is strictly smaller than that of the planted bisection. This in turn implies that the LP relaxation does not recover the planted bisection.

Before proceeding further, we introduce some notation that we will use to present our result. Consider a graph $G = (V, E)$, and let $\rho_G : V^2 \to \mathbb{Z}_{\geq 0} \cup \{+\infty\}$ denote the pairwise distances in G; i.e., $\rho_G(i, j)$ is the length of the shortest path from i to j in G, with $\rho_G(i, j) = +\infty$ if i and j belong to different connected components of G. Define the *diameter* and *average distance* in G as follows:

$$\rho_{\max}(G) := \max_{i,j \in V} \rho_G(i, j),$$
$$\rho_{\text{avg}}(G) := \mathbb{E}_{i,j \sim \text{Unif}(V)}[\rho_G(i, j)] = \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \rho_G(i, j).$$

Let us also give some intuition behind our argument, which will also give some explanation for the relevance of these quantities. Problem (LP-P) optimizes, as its name suggests, over a choice of a *metric* on the vertex set of G, and seeks to make the distances under this metric small between pairs of points that are adjacent in G. To find a non-integral feasible point with a large value, it is therefore natural to start with the metric ρ_G itself, and try to renormalize it to build a feasible x for Problem (LP-P). This is precisely what we will do in the proof of the next result; $\rho_{\text{avg}}(G)$ will be involved when we normalize x to have a fixed value of $\sum_{i<j} x_{ij}$, and $\rho_{\max}(G)$ when we adjust our construction to satisfy the other "triangle inequalities" included in the problem.

Theorem 5. *Let $G = (V, E)$ be a connected graph on $n \geq 5$ nodes. Define*

$$c(G) := \max \left\{ 0, \frac{3\rho_{\max}(G) - 4\rho_{\text{avg}}(G)}{1 - \frac{4}{n}} \right\}. \tag{17}$$

If

$$\frac{1 + c(G)}{2\rho_{\text{avg}}(G) + 2c(G)(1 - \frac{1}{n})} |E| < |E_0|, \tag{18}$$

then LP relaxation (LP-P) *does not recover the planted bisection.*

Proof. Define the point $\tilde{x} \in \mathbb{R}^{\binom{n}{2}}$ as

$$\tilde{x}_{ij} = \frac{\rho_G(i, j) + c(G)}{2\rho_{\text{avg}}(G) + 2c(G)(1 - \frac{1}{n})} = \frac{\rho_G(i, j) + c(G)}{\frac{4}{n^2} \sum_{1 \leq i < j \leq n} (\rho_G(i, j) + c(G))}.$$

We first show that $\tilde{x}$ is feasible for (LP-P). By construction $\tilde{x}$ satisfies the equality constraint (3). Also, since ρ_G is a metric and $c \geq 0$, $\tilde{x}$ satisfies inequalities (1). Thus it suffices to show that $\tilde{x}$ satisfies inequalities (2). To do this, we bound:

$$\tilde{x}_{ij} + \tilde{x}_{ik} + \tilde{x}_{jk} \leq \frac{3}{2} \cdot \frac{\rho_{\max}(G) + c(G)}{\rho_{\text{avg}}(G) + c(G)(1 - \frac{1}{n})}. \tag{19}$$

Two cases arise:

(i) $c(G) = 0$: in this case, we have $\rho_{\max}(G) \leq \frac{4}{3}\rho_{\text{avg}}(G)$ and hence we can further bound the right-hand side of inequality (19) as:

$$\tilde{x}_{ij} + \tilde{x}_{ik} + \tilde{x}_{jk} \leq \frac{3}{2} \cdot \frac{\rho_{\max}(G)}{\rho_{\text{avg}}(G)} \leq \frac{3}{2} \cdot \frac{4}{3} = 2.$$

(ii) $c(G) > 0$: in this case, we have $\rho_{\max}(G) > \frac{4}{3}\rho_{\text{avg}}(G)$ and $c(G) = (3\rho_{\max}(G) - 4\rho_{\text{avg}}(G))/(1 - \frac{4}{n})$; hence we can further bound the right-hand side of inequality (19) as:

$$\begin{aligned}
\tilde{x}_{ij} + \tilde{x}_{ik} + \tilde{x}_{jk} &\leq \frac{3}{2} \cdot \frac{\rho_{\max}(G)(1 - \frac{4}{n}) + 3\rho_{\max}(G) - 4\rho_{\text{avg}}(G)}{\rho_{\text{avg}}(G)(1 - \frac{4}{n}) + (3\rho_{\max}(G) - 4\rho_{\text{avg}}(G))(1 - \frac{1}{n})} \\
&= \frac{3}{2} \cdot \frac{4\rho_{\max}(G)(1 - \frac{1}{n}) - 4\rho_{\text{avg}}(G)}{3\rho_{\max}(G)(1 - \frac{1}{n}) - 3\rho_{\text{avg}}(G)} \\
&= 2.
\end{aligned}$$

Thus in either case $\tilde{x}$ satisfies inequalities (2), implying $\tilde{x}$ is feasible. It then follows that the objective value of (LP-P) at $\tilde{x}$, given by $\frac{1+c(G)}{\rho_{\text{avg}}(G)+c(G)(1-\frac{1}{n})} \cdot \frac{|E|}{2}$, provides an upper bound on its optimal value. Moreover, the objective value of (LP-P) corresponding to the planted bisection is equal to $|E_0|$. Hence, if condition (18) holds, the LP relaxation does not recover the planted bisection. □

The no-recovery condition (18) depends on $c(G)$ which in turn depends on the relative values of $\rho_{\text{avg}}(G)$ and $\rho_{\max}(G)$. Clearly, for any graph G we have $\rho_{\text{avg}}(G) \leq \rho_{\max}(G)$. As we detail in Section 4, for random graphs, very often these two quantities are quite close together, whereby $|E_0|$ must be quite small compared to $|E|$ in order for recovery to be possible. In short, this result lets us infer that, since typical distances between nodes in random graphs are comparable to the maximum such distance, the LP seldom succeeds in recovering the planted partition.

Before proceeding to those arguments, to illustrate the basic idea we explain how this analysis looks for a specific type of deterministic graph. Let $G = (V, E)$ be a graph such that $\rho_G(i, j) \leq 2$ for every $i, j \in V$, $\rho_G(i, j) = 2$ for some $i, j \in V$. That is, G is not the complete graph, but any pair of non-adjacent nodes in G have a common neighbor. In this case, we may explicitly compute $\rho_{\max}(G) = 2$ and

$$\rho_{\text{avg}}(G) = \frac{2}{n^2}\left(|E| + 2\left(\binom{n}{2} - |E|\right)\right) = 2\left(1 - \frac{1}{n}\right) - \frac{2}{n^2}|E|.$$

Thus $3\rho_{\max}(G) - 4\rho_{\text{avg}}(G) = -2 + \frac{8}{n} + \frac{8}{n^2}|E|$. If moreover n is sufficiently large and $|E| \leq \frac{n^2}{4} - n$, we have $c(G) = 0$. Thus the non-recovery condition (18) reduces to $|E_0|/|E| > \frac{1}{2\rho_{\text{avg}}(G)}$, where $\rho_{\text{avg}}(G) \in (1, 2)$. Therefore, in this type of graph our analysis implies that if a sufficiently large constant fraction of edges cross the planted bisection, then the LP cannot recover the planted bisection.

4 The stochastic block model

The *stochastic block model (SBM)* is a probability distribution over instances of the planted bisection problem, which in recent years has been studied extensively as a benchmark for community detection (see [1] for an extensive survey). The SBM can be seen as an extension of the *Erdős-Rényi (ER)* model for random graphs. In the ER model, edges are placed between each pair of nodes independently with probability p. In the SBM, there is an additional community structure built in to the random graph. In this paper, we focus on the simplest case, the *symmetric SBM on two communities*. In this model, for an even number n of nodes, we first choose a bisection V_1, V_2 of the set of nodes uniformly at random. We then draw edges between pairs of nodes $i, j \in V_1$ or $i, j \in V_2$ with probability p, and draw edges between pairs of nodes $i \in V_1$ and $j \in V_2$ with probability q. As in the ER model, all edges are chosen independently. In addition, we always assume $p > q$,

so that the average connectivity within each individual community is stronger than that between different communities.

We denote by $\mathcal{G}_{n,p}$ the ER model on n nodes with edge probability p, and by $\mathcal{G}_{n,p,q}$ the SBM on n nodes with edge probabilities p and q. We write $G \sim \mathcal{G}_{n,p}$ or $G \sim \mathcal{G}_{n,p,q}$ for G drawn from either distribution. Generally, we consider probability parameters depending on n, $p = p(n)$ and $q = q(n)$, but for the sake of brevity we often omit this dependence. When p and q are in fact constants independent of n, we refer to the graph $G \sim \mathcal{G}_{n,p,q}$ as *very dense*. When instead p and q scale as

$$p = \alpha n^{-\omega}, \qquad q = \beta n^{-\omega}, \tag{20}$$

where α, β and ω are parameters independent of n such that $\alpha > \beta > 0$ and $0 < \omega < 1$, we refer to the graph $G \sim \mathcal{G}_{n,p,q}$ as *dense* and finally when p and q scale as

$$p = \alpha \frac{\log n}{n}, \qquad q = \beta \frac{\log n}{n}, \tag{21}$$

where α and β are parameters independent of n such that $\alpha > \beta > 0$, we refer to the graph $G \sim \mathcal{G}_{n,p,q}$ as having *logarithmic degree*.

Let us briefly review some of the existing literature on recovery in the SBM, to establish a baseline for our results. In [2], the authors show that the minimum bisection problem recovers the planted bisection with high probability for very dense and dense graphs as long as $p > q$. For graphs of logarithmic degree, they show that the minimum bisection problem recovers the planted bisection with high probability if and only if $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$. This gives an "information-theoretic limit" on recovery, in the sense that if $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}$, then no algorithm (efficient or not) can recover the planted bisection with high probability. Moreover, the authors consider an SDP relaxation of the minimum bisection problem, and conjecture that in the logarithmic regime the SDP relaxation recovers the planted bisection whenever the minimum bisection problem does. The authors of [2] provided a proof of a partial result in this direction, and the subsequent works [23, 5] gave complete proofs. In Section 5.3 we show that the deterministic recovery condition in [5] implies that the SDP relaxation also recovers the planted bisection for very dense and dense graphs up to the information theoretic threshold. The final picture emerging from this analysis is that the SDP relaxation is "as good as" the information-theoretically optimal integer program.

In this section, we investigate the theoretical performance of the LP relaxation for recovery in the SBM. We show that in any of the three regimes described above, the LP relaxation does *not* recover the planted bisection with high probability in the full regions of parameters where it is information-theoretically possible to do so. More specifically, we first obtain a sufficient condition under which the LP relaxation recovers the planted bisection for very dense graphs. Subsequently, we obtain non-recovery conditions for very dense and dense graphs. Finally we show that for graphs of logarithmic degree, the LP relaxation fails to recover the planted bisection with high probability. In short, the LP relaxation is not "as good as" the SDP relaxation in recovering the communities in the SBM.

4.1 Recovery in SBM

In this section, we consider very dense random graphs and obtain a sufficient condition under which the LP relaxation recovers the planted bisection with high probability. To this end, we first show that in the very dense regime, the regularity assumptions of Theorem 4 are essentially not restrictive. Subsequently, we employ Theorem 4 to obtain a recovery guarantee for the very dense SBM. The following propositions provide sufficient conditions under which bipartite ER graphs and general ER graphs are contained in or contain regular graphs of comparable average degrees. In the remainder of the paper, for a sequence of events A_n depending on n, we say A_n occurs *with high probability* if $\mathbb{P}[A_n] \to 1$ as $n \to \infty$.

Proposition 3. *There exists a constant $C_{\mathrm{reg},2} > 0$ such that, for any $q \in (0,1)$, with high probability $G \sim \mathcal{G}_{n,0,q}$ is a subgraph of a d_{reg}-regular bipartite graph with the same bipartition, where $d_{\mathrm{reg}} = \frac{qn}{2} + C_{\mathrm{reg},2}\sqrt{\frac{n}{2}\log\frac{n}{2}}$.*

Proposition 4. *There exists a constant $C_{\mathrm{reg},1} > 0$ such that, for any $p \in (0,1)$, with high probability $G \sim \mathcal{G}_{n,p}$ contains as a subgraph a d_{reg}-regular graph on the same node set, where $d_{\mathrm{reg}} = pn - C_{\mathrm{reg},1}\sqrt{n\log n}$.*

We note that $\frac{qn}{2}$ and pn are the expected degrees of any given vertex in $G \sim \mathcal{G}_{n,0,q}$ and $G \sim \mathcal{G}_{n,p}$, respectively. The proofs of Propositions 3 and 4 are given in Sections 5.1 and 5.2, respectively, and consist of performing a probabilistic analysis of conditions for the existence of "graph factors" as developed in classical results of graph theory literature [40]. As such, these results are of independent interest to both optimization and applied probability communities.

By combining Theorem 4 and Propositions 4 and 3, we obtain a sufficient condition for recovery in the very dense regime.

Theorem 6. *Let $0 < q < p < 1$ with $p - q > \frac{1}{2}$. Then, the LP relaxation recovers the planted bisection in $G \sim \mathcal{G}_{n,p,q}$ with high probability.*

Proof. As before, denote by G_1 and G_2 the subgraphs of G induced by V_1 and V_2, respectively, and denote by G_0 the bipartite subgraph of G of all edges between V_1 and V_2. We start by regularizing the bipartite graph G_0. By Proposition 3, with high probability, it is possible to add edges to G_0 to make it d_{out}-regular, for $d_{\text{out}} = \frac{qn}{2} + C_{\text{reg},2}\sqrt{\frac{n}{2}\log(\frac{n}{2})}$ and $C_{\text{reg},2}$ a universal constant. Similarly, by Proposition 4, with high probability, it is possible to remove edges from G_1 and G_2 to make them d_{in}-regular, for $d_{\text{in}} = \frac{pn}{2} - C_{\text{reg},1}\sqrt{\frac{n}{2}\log(\frac{n}{2})}$ and $C_{\text{reg},1}$ another universal constant. From Theorem 4 it follows that the LP relaxation recovers the planted bisection with high probability, if $d_{\text{in}} - d_{\text{out}} \geq \frac{n}{4}$. Substituting for the values of d_{in} and d_{out} specified above gives that, with high probability, it suffices to have

$$\frac{pn}{2} - \frac{qn}{2} - (C_{\text{reg},1} + C_{\text{reg},2})\sqrt{\frac{n}{2}\log\left(\frac{n}{2}\right)} \geq \frac{n}{4}.$$

Dividing both side of the above inequality by n and letting $n \to \infty$ then gives the result, since $p - q > \frac{1}{2}$ by assumption. □

To better understand the tightness of the sufficient condition of Theorem 6, we conduct a simulation, depicted in Figure 2. As can be seen in this figure, as an artifact of our proof technique, the condition $p - q > \frac{1}{2}$ is somewhat suboptimal. While a better dual certificate construction scheme may improve this condition, as we can see from this figure, such an improvement is not going to be significant.

4.2 Non-recovery in SBM

We now consider random graphs in various regimes and provide conditions under which the recovery in these regimes is not possible. Our basic technique is to evaluate both sides of the condition (18) in Theorem 5 asymptotically for the SBM, and to verify that in suitable regimes, the condition holds with high probability. To this end, for a graph $G \sim \mathcal{G}_{n,p,q}$, we need to bound $c(G)$ defined by (17). This in turn amounts to bounding the quantities $\rho_{\max}(G)$ and $\rho_{\text{avg}}(G)$ in different regimes. Luckily the distance distributions of ER graphs have been studied extensively in the literature. Utilizing the existing work on ER graphs, we obtain similar results for the SBM. We refer the reader to Section 5.4 for statements and proofs.

Our first result gives a collection of non-recovery conditions for very dense and dense graphs.

Theorem 7. *Let $p = \alpha n^{-\omega}$ and $q = \beta n^{-\omega}$ for $\alpha, \beta > 0$ and $\omega \in [0, 1)$, with $\alpha, \beta \in (0, 1)$ if $\omega = 0$. Suppose that one of the following conditions holds:*

1. *$\omega = 0$ and*
$$\beta > \max\left\{\frac{3 - \alpha - \sqrt{(3-\alpha)^2 - 4\alpha}}{2},\ 2\alpha - 1\right\}. \tag{22}$$
2. *$\omega \in (0, 1)$ with $\frac{1}{1-\omega} \notin \mathbb{N}$ and*
$$\beta > \frac{1}{2\lceil \frac{1}{1-\omega} \rceil - 1} \cdot \alpha. \tag{23}$$
3. *$\omega = \frac{1}{2}$ and*
$$\beta > \max\left\{\frac{1 - 2e^{-\alpha^2}}{2 - e^{-\alpha^2}},\ \frac{1}{3 + 2e^{-\alpha^2}}\right\} \cdot \alpha. \tag{24}$$

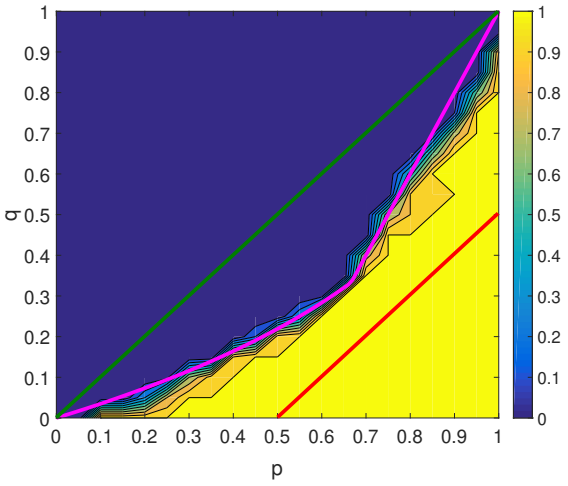

Figure 2: The empirical probability of success of the LP relaxation in recovering the planted bisection in the very dense regime. For each fixed pair (p, q), we fix $n = 100$ and the number of trials to be 20. Then, for each fixed (p, q), we count the number of times the LP relaxation recovers the planted bisection. Dividing by the number of trials, we obtain the empirical probability of success. In red, we plot the threshold for recovery of the LP as given by Theorem 6. In magenta, we plot the threshold for failure of the LP relaxation as given by Part 1 of Theorem 7. In green, we plot the information-theoretic limit for recovery in the very dense regime.

4. $\omega \in (0, 1)$ *with* $\frac{1}{1-\omega} \in \mathbb{N} \setminus \{2\}$ *and*

$$\beta > \frac{1}{2 \cdot \frac{1}{1-\omega} - 1 + 2\exp(-\alpha^{\frac{1}{1-\omega}})} \cdot \alpha. \tag{25}$$

Then, with high probability, the LP fails to recover the planted bisection.

Proof. First, we bound the fraction of edges across the bisection in the SBM. Let $p, q = \Omega(\log n/n)$ and having $\lim_{n\to\infty} p/q$ existing and taking a value in $(0, 1]$. Then, for any $\epsilon > 0$, an application of Hoeffding's inequality to both $|E|$ and $|E_0|$ shows that

$$\lim_{n\to\infty} \mathbb{P}\left[(1-\epsilon)\frac{q}{p+q} \le \frac{|E_0|}{|E|} \le (1+\epsilon)\frac{q}{p+q}\right] = 1, \tag{26}$$

where we note that $\frac{\mathbb{E}|E_0|}{\mathbb{E}|E|} = \frac{q}{p+q} + o(1)$, explaining the appearance of this quantity.

Next, define

$$b(G) := \frac{1 + c(G)}{2\rho_{\text{avg}}(G) + 2c(G)(1 - \frac{1}{n})}. \tag{27}$$

By Theorem 5, if $|E_0|/|E| > b(G)$, the LP does not recover the planted bisection. Thus our task consists in obtaining an upper bound on $b(G)$. First, notice that $b(G)$ is increasing in $c(G)$ since $\rho_{\text{avg}}(G) \le \frac{2}{n^2} \cdot \binom{n}{2} = 1 - \frac{1}{n}$. Moreover, $b(G)$ is decreasing in $\rho_{\text{avg}}(G)$. Hence, to upper bound $b(G)$ it suffices to upper bound $\rho_{\max}(G)$ and to lower bound $\rho_{\text{avg}}(G)$. We will treat the four cases individually.

Condition 1: $\omega = 0$. By Proposition 16, for any $\epsilon > 0$, the distance distribution in G satisfies, with high probability,

$$(1-\epsilon)\left(2 - \frac{p+q}{2}\right) \le \rho_{\text{avg}}(G) \le \rho_{\max}(G) = 2. \tag{28}$$

On the event that this holds, we have

$$3\rho_{\max}(G) - 4\rho_{\text{avg}}(G) \le 6 - 4(1-\epsilon)\left(2 - \frac{p+q}{2}\right) \le 2(p+q-1) + 8\epsilon.$$

Therefore, we have, for sufficiently small ϵ, with high probability,

$$c(G) \le \frac{1}{1-\frac{4}{n}} \left(2\max\{0, p+q-1\} + 8\epsilon\right) \le 2\max\{0, p+q-1\} + 9\epsilon.$$

We now consider two cases depending on the output of the maximum in the first term.

Case 1.1: $p+q \le 1$. In this case, with high probability $c(G) \le 9\epsilon$. By (28), we also have, with high probability, $\rho_{\text{avg}}(G) \ge (1-\epsilon)\frac{4-p-q}{2}$. Thus from the definition of $b(G)$, for ϵ sufficiently small, with high probability,

$$b(G) \le (1+O(\epsilon))\frac{1}{4-p-q}.$$

Therefore, by (18) and (26) and by taking ϵ sufficiently small, it suffices to have

$$\frac{1}{4-p-q} < \frac{q}{p+q}.$$

The above is equivalent to $q > \frac{1}{2}(3-p-\sqrt{(3-p)^2-4p})$ upon solving the inequality for q, completing the proof.

Case 1.2: $p+q > 1$. In this case, with high probability $c(G) \le 2(p+q-1)+9\epsilon$. For ϵ sufficiently small, with high probability,

$$b(G) \le (1+O(\epsilon))\frac{1+2(p+q-1)}{4-p-q+4(p+q-1)} = (1+O(\epsilon))\frac{2p+2q-1}{3p+3q}.$$

As in Case 1.1, it then suffices to have

$$\frac{2p+2q-1}{3p+3q} < \frac{q}{p+q}.$$

The above is equivalent to $q > 2p-1$ upon solving the inequality for q, completing the proof.

Condition 2: $\omega \in (0,1)$ *with* $\frac{1}{1-\omega} \notin \mathbb{N}$. By Proposition 17, in this regime for any $\epsilon > 0$, the distance distribution in G with high probability satisfies

$$(1-\epsilon)\left\lceil \frac{1}{1-\omega} \right\rceil \le \rho_{\text{avg}}(G) \le \rho_{\max}(G) = \left\lceil \frac{1}{1-\omega} \right\rceil.$$

On this event, we have $3\rho_{\max}(G) - 4\rho_{\text{avg}}(G) \le -\lceil \frac{1}{1-\omega} \rceil + O(\epsilon)$, so for sufficiently small ϵ this quantity is negative, whereby $c(G) = 0$. In this case, we have

$$b(G) = \frac{1}{2\rho_{\text{avg}}(G)} \le (1+O(\epsilon))\frac{1}{2\lceil \frac{1}{1-\omega} \rceil},$$

and thus it suffices to have

$$\frac{\beta}{\alpha+\beta} > \frac{1}{2\lceil \frac{1}{1-\omega} \rceil}.$$

The above condition is equivalent to condition (23) after solving the inequality for β, completing the proof.

Condition 3: $\omega = \frac{1}{2}$. In this case, $\frac{1}{1-\omega} = 2$. By Proposition 18, in this regime for any $\epsilon > 0$, the distance distribution in G with high probability satisfies

$$(1-\epsilon)(2+\exp(-\alpha^2)) \le \rho_{\text{avg}}(G) \le \rho_{\max}(G) = 3.$$

On this event, we have

$$3\rho_{\max}(G) - 4\rho_{\text{avg}}(G) \le 9 - 4(1-\epsilon)(2+\exp(-\alpha^2)) \le 1 - 4\exp(-\alpha^2) + O(\epsilon).$$

We then consider two cases:

Case 3.1: $\exp(-\alpha^2) \geq \frac{1}{4}$. In this case, we have $3\rho_{\max}(G) - 4\rho_{\text{avg}}(G) = O(\epsilon)$, whereby $c(G) = O(\epsilon)$ with high probability as well. Thus,

$$b(G) \leq (1 + O(\epsilon))\frac{1}{2\rho_{\text{avg}}(G)} \leq (1 + O(\epsilon))\frac{1}{4 + 2\exp(-\alpha^2)}.$$

It then suffices to have

$$\frac{\beta}{\alpha + \beta} > \frac{1}{4 + 2\exp(-\alpha^2)},$$

which, upon solving the inequality for β is equivalent to $\beta > \frac{1}{3+2\exp(-\alpha^2)} \cdot \alpha$.

Case 3.2: $\exp(-\alpha^2) < \frac{1}{4}$. In this case, we have $c(G) \leq (1 + O(\epsilon))(1 - 4\exp(-\alpha^2))$ with high probability. Thus,

$$b(G) \leq (1 + O(\epsilon))\frac{1 + (1 - 4\exp(-\alpha^2))}{2\rho_{\text{avg}}(G) + 2(1 - 4\exp(-\alpha^2))} \leq (1 + O(\epsilon))\frac{1 - 2\exp(-\alpha^2)}{3 - 3\exp(-\alpha^2)}.$$

It then suffices to have

$$\frac{\beta}{\alpha + \beta} > \frac{1 - 2\exp(-\alpha^2)}{3 - 3\exp(-\alpha^2)},$$

which, upon solving the inequality for β is equivalent to $\beta > \frac{1-2\exp(-\alpha^2)}{2-\exp(-\alpha^2)} \cdot \alpha$.

Condition 4: $\omega \in (0, 1)$ *with* $\frac{1}{1-\omega} \in \mathbb{N} \setminus \{2\}$. In this case, $\omega = \frac{k-1}{k}$ for some $k \in \mathbb{N}$ with $k \geq 3$, and $\frac{1}{1-\omega} = k$. By Proposition 18, in this regime for any $\epsilon > 0$, the distance distribution in G with high probability satisfies

$$(1 - \epsilon)(k + \exp(-\alpha^k)) \leq \rho_{\text{avg}}(G) \leq \rho_{\max}(G) = k + 1.$$

On this event, we have

$$3\rho_{\max}(G) - 4\rho_{\text{avg}}(G) \leq 3 - (1 - 4\epsilon)k - 4(1 - \epsilon)\exp(-\alpha^k) \leq 12\epsilon - 4(1 - \epsilon)\exp(-\alpha^k),$$

where the second inequality is valid since $k \geq 3$. It then follows that for ϵ sufficiently small, $3\rho_{\max}(G) - 4\rho_{\text{avg}}(G) < 0$, and hence for such ϵ with high probability $c(G) = 0$. On this event, we have

$$b(G) = \frac{1}{2\rho_{\text{avg}}(G)} \leq (1 + O(\epsilon))\frac{1}{2k + 2\exp(-\alpha^k)},$$

and thus it suffices to check that

$$\frac{\beta}{\alpha + \beta} > \frac{1}{2k + 2\exp(-\alpha^k)}.$$

The above inequality is equivalent to inequality (25) after solving the inequality for β, completing the proof. □

The non-recovery thresholds of Theorem 7 are plotted in Figure 4.2 for several values of ω. Moreover, condition (22) is plotted in Figure 2 as well. As can be seen from this figure, our non-recovery threshold in the very dense regime is almost in perfect agreement with our empirical observations.

Finally, we show that the LP relaxation fails to recover the planted bisection in the logarithmic regime.

Theorem 8. *Suppose* $p = \alpha\frac{\log n}{n}$ *and* $q = \beta\frac{\log n}{n}$ *with* $\alpha, \beta > 0$. *Then, with high probability, the LP relaxation fails to recover the planted bisection in* $G \sim \mathcal{G}_{n,p,q}$.

Proof. If $\frac{\alpha+\beta}{2} \leq 1$, then the statement follows from the information-theoretic bound of [2]. Let us suppose that $\frac{\alpha+\beta}{2} > 1$; in this case, since $\frac{q}{p+q} = \frac{\beta}{\alpha+\beta} > 0$ is a constant, it suffices to show that for any $\delta > 0$, with high probability, $b(G) < \delta$. By Proposition 19, in the logarithmic regime, for any $\epsilon > 0$, the distance distribution in G satisfies, with high probability,

$$(1 - \epsilon)\frac{\log n}{\log\log n} \leq \rho_{\text{avg}}(G) \leq \rho_{\max}(G) \leq (1 + \epsilon)\frac{\log n}{\log\log n}. \tag{29}$$

It then follows that with high probability $c(G) = 0$ and $b(G) \leq \frac{1}{2(1-\epsilon)}\frac{\log\log n}{\log n}$ and hence the result follows. □

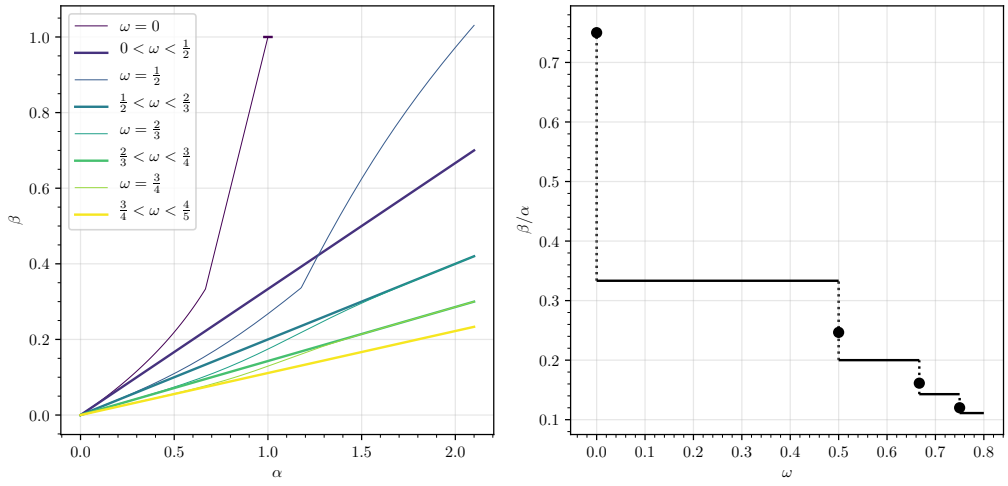

Figure 3: Non-recovery thresholds as given by Theorem 7 for SBM in the very dense and dense regimes. For each choice of ω, our results prove that with high probability the LP does not recover the planted bisection for all choices of (α, β) lying above the given curve in the left panel. The curve for $\omega = 0$ is identical to that compared with numerical results in Figure 2. In the right panel, we show the value of β/α at the non-recovery threshold value of β for a fixed $\alpha = 0.8$ as ω varies. In contrast, the same ratio for the SDP relaxation is always equal to 1.

In [28], the authors consider the metric relaxation of the sparsest cut problem. They show that, for constant degree expander graphs, which includes random regular graphs of constant degree, the LP relaxation has an integrality gap of $\Theta(\log n)$. We outline this argument: a d-regular expander satisfies, by definition, $|E(U, U^c)| \gtrsim d \min\{|U|, |U^c|\}$ for all $U \subseteq V$, with "$\gtrsim$" hiding a constant not depending on n. Thus the sparsest cut objective function is $\min_U \frac{|E(U,U^c)|}{|U| \cdot |U^c|} \gtrsim \frac{d}{n}$. On the other hand, the authors of [28] show that the value of the metric relaxation is $\lesssim \frac{\log d}{\log n} \cdot \frac{d}{n}$. For d constant, this shows the integrality gap of $\Theta(\log n)$.

For random regular and ER graphs of average degree $d = d(n) \gtrsim \log n$, the same lower and upper bounds for the sparsest cut and its metric relaxation are valid. Up to an adjustment of constants, we also expect them to hold for SBM graphs, since these are nothing but "non-uniform ER graphs". For $d \sim \log n$, we have $\frac{\log d}{\log n} \sim \frac{\log \log n}{\log n} = o(1)$, so we expect the metric relaxation to have a diverging integrality gap over all SBM graphs. Thus it seems plausible that the proof techniques of [28] could reproduce the non-recovery result of Theorem 8. On the other hand, for $d \sim n^{1-\omega}$ as in our dense regimes, we have $\frac{\log d}{\log n} \sim 1 - \omega$ is a constant. Since we expect that working with SBM graphs rather than ER graphs also creates a constant order adjustment in the integrality gap, the argument of [28] does not show non-recovery, and the more refined analysis of Theorem 7 is necessary.

Our results do still leave open the question of whether the LP relaxation fails to recover the planted bisection in the entirety of the dense regimes $p, q \sim n^{-\omega}$ with $\omega \in (0, 1)$ of the SBM, or whether they undergo a transition from recovery to non-recovery as in the very dense regime $\omega = 0$. As we detailed at the end of Section 3.1.1, it may be possible to strengthen our deterministic recovery condition in Proposition 1 by characterizing the input graph in terms of the spectrum of its adjacency matrix rather than the two parameters $d_{\text{in}}, d_{\text{out}}$. This could imply that recovery occurs with high probability in part of the dense regime; a full characterization of the regime over which the LP relaxation undergoes a transition from recovery to non-recovery is a subject for future research. Finally we must point out that triangle inequalities constitute a very small fraction of facet-defining inequalities for the cut polytope and many more classes of facet-defining inequalities are known for this polytope [19]. It would be interesting to investigate the recovery properties of stronger LP relaxations for the min-bisection problem.

5 Technical proofs

5.1 Proof of Proposition 3

To prove the statement, we make use of two results. The first result provides a necessary and sufficient condition under which a bipartite graph has a b-factor. Consider a graph $G = (V, E)$; given a vector $b \in \mathbb{Z}_{\geq 0}^V$, a *$b$-factor* is a subset F of E such that b_v edges in F are incident to v for all $v \in V$. For notational simplicity, given a vector $b \in \mathbb{Z}_{\geq 0}^V$ and a subset $U \subseteq V$, we define $b(U) := \sum_{v \in U} b_v$. Moreover, for $U, U' \subseteq V$, we denote by $N(U, U')$ the number of edges in G with one endpoint in U and one endpoint in U'.

Proposition 5 (Corollary 21.4a in [37])**.** *Let $G = (V, E)$ be a bipartite graph and let $b \in \mathbb{Z}_{\geq 0}^V$. Then G has a b-factor if and only if, for each $U \subseteq V$,*

$$N(U, U) \geq b(U) - \frac{1}{2} b(V).$$

The second result that we need is concerned with an upper bound on the maximum node degree in a bipartite dense random graph.

Proposition 6. *Let $G \sim \mathcal{G}_{2m,0,q}$ for some $q \in (0, 1)$. Denote by $d_{\max}$ the maximum node degree of G. Then, there exists a constant $C > 0$ such that, with high probability, $d_{\max} \leq qm + C\sqrt{m \log m}$.*

Proof. We introduce the bipartite adjacency matrix of G: let $X \in \mathbb{R}^{V_1 \times V_2}$, where when $i \in V_1$ and $j \in V_2$, then $X_{ij} = 1$ if i and j are adjacent in G and $X_{ij} = 0$ otherwise. Then, when $G \sim \mathcal{G}_{2m,0,q}$, X_{ij} has i.i.d. entries equal to 1 with probability q and 0 otherwise. In particular, $\mathbb{E}[X_{ij}] = q$.

The degree of $i \in V_1$ is given by $d(i) = \sum_{j \in V_2} X_{ij}$. In particular, the $d(i)$ are themselves i.i.d. random variables. Since $X_{ij} \in [0, 1]$, Hoeffding's inequality applies to each $d(i)$, giving, for $C > 0$ a large constant to be fixed later,

$$\begin{aligned}
\mathbb{P}\left[d(i) > qm + C\sqrt{m \log m}\right] &= \mathbb{P}\left[\sum_{j \in V_2} (X_{ij} - \mathbb{E}[X_{ij}]) > C\sqrt{m \log m}\right] \\
&\leq \exp\left(\frac{-2C^2 m \log m}{m}\right) = m^{-2C^2}.
\end{aligned}$$

Since each $d(i)$ is independent for distinct $i \in V_1$, we moreover have, for some fixed $i_1 \in V_1$,

$$\begin{aligned}
&\mathbb{P}\left[d(i) > qm + C\sqrt{m \log m} \text{ for some } i \in V_1\right] \\
&\qquad = 1 - \mathbb{P}\left[d(i) \leq qm + C\sqrt{m \log m} \text{ for all } i \in V_1\right] \\
&\qquad = 1 - \left(1 - \mathbb{P}\left[d(i_1) > qm + C\sqrt{m \log m}\right]\right)^m \\
&\qquad \leq 1 - \left(1 - m^{-2C^2}\right)^m = m^{-2C^2} \left(\sum_{k=0}^{m-1} \left(1 - m^{-2C^2}\right)^k\right) \\
&\qquad \leq m \cdot m^{-2C^2} = m^{1-2C^2}.
\end{aligned}$$

Symmetrically, the same bound applies to V_1 replaced with V_2, so we find

$$\mathbb{P}\left[d_{\max} > qm + C\sqrt{m \log m}\right] \leq 2m^{1-2C^2},$$

and setting C large enough gives the result. □

We can now proceed with the proof of Proposition 3. We first set some useful notation. Let V be the node set of G and let V_1, V_2 be the bipartition of V, so that $|V_1| = |V_2| = m$. For $U \subseteq V$, let $U_1 = U \cap V_1$ and $U_2 = U \cap V_2$, whereby U is the disjoint union of U_1 and U_2. For $v \in V$, let $d(v)$ be the degree of v.

Let $\bar{G}$ be the bipartite graph complement of G, with the same node partition. Finally, for $U, U' \subseteq V$, let $\bar{N}(U, U')$ be the number of edges in $\bar{G}$ with one endpoint in U and one endpoint in U'.

The statement of the theorem is equivalent to $\bar{G}$ having a b-factor for

$$b(v) = d_{\text{reg}} - d(v).$$

Let $d_{\max}$ be the maximum node degree of G. By Proposition 6, for sufficiently large $C_{\text{reg},2}$, we will have $d_{\text{reg}} \geq d_{\max}$ with high probability, and thus $b(v) \geq 0$ with high probability. Let $A_{\deg}$ be the event that this occurs.

On the event $A_{\deg}$, by Proposition 5, such a b-factor exists if and only if, for each $U \subseteq V$,

$$\bar{N}(U, U) \geq b(U) - \frac{1}{2} b(V). \tag{30}$$

Let A_U be the event that this occurs. Then, whenever the event $A = A_{\deg} \cap \bigcap_{U \subseteq V} A_U$ occurs, G is a subgraph of a d_{reg}-regular bipartite graph. Thus it suffices to show that $\mathbb{P}[A^c] \to 0$ as $n \to \infty$. Taking a union bound,

$$\mathbb{P}[A^c] \leq \mathbb{P}[A^c_{\deg}] + \sum_{U \subseteq V} \mathbb{P}[A^c_U]. \tag{31}$$

We have already observed that for $C_{\text{reg},2}$ large enough, the first summand tends to zero, so it suffices to control the remaining sum.

Let us first rewrite each side of the equation (30). For any $U \subseteq V$, we have

$$b(U) = d_{\text{reg}}|U| - \sum_{u \in U} d(u) = d_{\text{reg}}|U| - 2N(U, U) - N(U, U^c),$$

and therefore

$$b(U) - \frac{1}{2} b(V) = d_{\text{reg}} \left(|U| - m\right) - N(U, U) + N(U^c, U^c).$$

Also,

$$\bar{N}(U, U) = |U_1| \cdot |U_2| - N(U, U).$$

Thus,

$$A_U = \left\{ \bar{N}(U, U) \geq b(U) - \frac{1}{2} b(V) \right\} = \left\{ N(U^c, U^c) \leq |U_1| \cdot |U_2| + d_{\text{reg}} \left(m - |U|\right) \right\}.$$

Note that

$$N(U^c, U^c) \leq |V_1 \setminus U_1| \cdot |V_2 \setminus U_2| = \left(m - |U_1|\right)\left(m - |U_2|\right) = |U_1| \cdot |U_2| - m\left(|U| - m\right).$$

Since, for sufficiently large m, $d_{\text{reg}} \leq m$, if $|U| \geq m$ then A_U always occurs, so in fact for such m we have $\mathbb{P}[A^c_U] = 0$ whenever $|U| \geq m$. Thus,

$$\lim_{m \to \infty} \sum_{U \subseteq V} \mathbb{P}[A^c_U] = \lim_{m \to \infty} \sum_{\substack{U \subseteq V \\ |U| < m}} \mathbb{P}[A^c_U]. \tag{32}$$

Recall that we set $X \in \mathbb{R}^{V_1 \times V_2}$ to be the bipartite adjacency matrix of G. When $G \sim \mathcal{G}_{n,0,q}$, X_{ij} has i.i.d. entries equal to 1 with probability q and 0 otherwise. In particular, $\mathbb{E}[X_{ij}] = q$. Also, we may compute edge counts as $N(U, U) = \sum_{i \in U_1} \sum_{j \in U_2} X_{ij}$.

Therefore, we may rewrite

$$\begin{aligned}
A_U &= \left\{ \sum_{i\in V_1\setminus U_1} \sum_{j\in V_2\setminus U_2} X_{ij} \le |U_1|\cdot|U_2| + d_{\mathrm{reg}}(m-|U|) \right\} \\
&= \left\{ \sum_{i\in V_1\setminus U_1} \sum_{j\in V_2\setminus U_2} (X_{ij}-\mathbb{E}[X_{ij}]) \le |U_1|\cdot|U_2| - q(m-|U_1|)(m-|U_2|) + d_{\mathrm{reg}}(m-|U|) \right\} \\
&= \left\{ \sum_{i\in V_1\setminus U_1} \sum_{j\in V_2\setminus U_2} (X_{ij}-\mathbb{E}[X_{ij}]) \le (1-q)|U_1|\cdot|U_2| + (d_{\mathrm{reg}}-qm)(m-|U|) \right\} \\
&= \left\{ \sum_{i\in V_1\setminus U_1} \sum_{j\in V_2\setminus U_2} (X_{ij}-\mathbb{E}[X_{ij}]) \le (1-q)|U_1|\cdot|U_2| + C_{\mathrm{reg},2}\sqrt{m\log m}(m-|U|) \right\}.
\end{aligned}$$

In this form, since $X_{ij}\in[0,1]$, Hoeffding's inequality (Theorem 2.2.6 in [41]) applies and gives

$$\begin{aligned}
\mathbb{P}[A_U^c] &= \mathbb{P}\left[\sum_{i\in V_1\setminus U_1} \sum_{j\in V_2\setminus U_2} (X_{ij}-\mathbb{E}[X_{ij}]) > (1-q)|U_1|\cdot|U_2| + C_{\mathrm{reg},2}\sqrt{m\log m}(m-|U|) \right] \\
&\le \exp\left(-\frac{2\big((1-q)|U_1|\cdot|U_2| + C_{\mathrm{reg},2}\sqrt{m\log m}(m-|U|)\big)^2}{(m-|U_1|)(m-|U_2|)} \right)
\end{aligned}$$

Note further that this bound, for a given U, depends only on $|U_1|$ and $|U_2|$. Thus, in the sum appearing in (32), we may group terms according to new scalar variables $a=|U_1|$ and $b=|U_2|$ and bound

$$\begin{aligned}
\sum_{\substack{U\subseteq V\\ |U|<m}} \mathbb{P}[A_U^c] &= \sum_{\substack{a,b\in[m]\\ a+b<m}} \sum_{\substack{U_1\in\binom{V_1}{a}\\ U_2\in\binom{V_2}{b}}} \mathbb{P}[A_{U_1\cup U_2}^c] \\
&\le \sum_{\substack{a,b\in[m]\\ a+b<m}} \sum_{\substack{U_1\in\binom{V_1}{a}\\ U_2\in\binom{V_2}{b}}} \exp\left(-\frac{2\big((1-q)ab + C_{\mathrm{reg},2}\sqrt{m\log m}(m-a-b)\big)^2}{(m-a)(m-b)} \right) \\
&= \sum_{\substack{a,b\in[m]\\ a+b<m}} \binom{m}{a}\binom{m}{b} \exp\left(-\frac{2\big((1-q)ab + C_{\mathrm{reg},2}\sqrt{m\log m}(m-a-b)\big)^2}{(m-a)(m-b)} \right). \qquad (33)
\end{aligned}$$

Now, note that since $(m-a)(m-b)=ab+m(m-a-b)$, for any choice of $a,b\in[m]$ with $a+b<m$ we have either $ab\ge\frac{1}{2}(m-a)(m-b)$ or $m(m-a-b)\ge\frac{1}{2}(m-a)(m-b)$. Note also that the latter is equivalent to $m-a-b\ge\frac{1}{2m}(m-a)(m-b)$. Notating this decomposition more formally, define

$$\begin{aligned}
\mathcal{I} &= \{(a,b)\in[m]^2 : a+b<m\} \\
\mathcal{I}_1 &= \left\{ (a,b)\in[m]^2 : a+b<m, ab\ge\frac{1}{2}(m-a)(m-b) \right\} \\
\mathcal{I}_2 &= \left\{ (a,b)\in[m]^2 : a+b<m, m-a-b\ge\frac{1}{2m}(m-a)(m-b) \right\}.
\end{aligned}$$

Then, the above observation says that $\mathcal{I}=\mathcal{I}_1\cup\mathcal{I}_2$.

When $(a,b)\in\mathcal{I}_1$, then we have

$$\begin{aligned}
\frac{2\big((1-q)ab + C_{\mathrm{reg},2}\sqrt{m\log m}(m-a-b)\big)^2}{(m-a)(m-b)} &\ge \frac{2(1-q)^2(ab)^2}{(m-a)(m-b)} \\
&\ge \frac{1}{2}(1-q)^2(m-a)(m-b), \qquad (34)
\end{aligned}$$

and when $(a,b) \in \mathcal{I}_2$, then we have

$$\frac{2\big((1-q)ab + C_{\mathrm{reg},2}\sqrt{m\log m}(m-a-b)\big)^2}{(m-a)(m-b)} \geq \frac{2C_{\mathrm{reg},2}^2 m\log m(m-a-b)^2}{(m-a)(m-b)}$$
$$\geq \frac{1}{2}C_{\mathrm{reg},2}^2\frac{\log m}{m}(m-a)(m-b). \tag{35}$$

For sufficiently large m, the right-hand side of (35) is always smaller than the right-hand side of (34) for all $(a,b) \in \mathcal{I}$. On the other hand, either (34) or (35) holds for any $(a,b) \in \mathcal{I}$, since $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$. Thus for sufficiently large m, for all $(a,b) \in \mathcal{I}$, we have

$$\frac{2\big((1-q)ab + C_{\mathrm{reg},2}\sqrt{m\log m}(m-a-b)\big)^2}{(m-a)(m-b)} \geq \frac{1}{2}C_{\mathrm{reg},2}^2\frac{\log m}{m}(m-a)(m-b).$$

Using this in (33), we may bound

$$\lim_{m\to\infty}\sum_{\substack{U\subseteq V\\|U|<m}}\mathbb{P}[A_U^c] \leq \lim_{m\to\infty}\sum_{\substack{a,b\in[m]\\a+b<m}}\binom{m}{a}\binom{m}{b}\exp\left(-\frac{1}{2}C_{\mathrm{reg},2}^2\frac{\log m}{m}(m-a)(m-b)\right).$$

Whenever $a+b<m$, then either $a<m/2$ or $b<m/2$, so symmetrizing the sum we have

$$\leq \lim_{m\to\infty} 2\sum_{\substack{a\in[m/2]\\b\in[m]\\a+b<m}}\binom{m}{a}\binom{m}{b}\exp\left(-\frac{1}{2}C_{\mathrm{reg},2}^2\frac{\log m}{m}(m-a)(m-b)\right)$$
$$\leq \lim_{m\to\infty} 2\sum_{\substack{a\in[m/2]\\b\in[m]\\a+b<m}}\binom{m}{a}\binom{m}{b}\exp\left(-\frac{1}{4}C_{\mathrm{reg},2}^2\log m\cdot(m-b)\right).$$

Now, note that $\binom{m}{b} = \binom{m}{m-b} \leq m^{m-b} = \exp(\log m\cdot(m-b))$, and $\binom{m}{a} \leq m^a = \exp(\log m\cdot a)$. Moreover, $a < m-b$ in all terms in the summation, so $\binom{m}{a}\binom{m}{b} \leq \exp(2\log m\cdot(m-b))$. Therefore, we continue

$$\leq \lim_{m\to\infty} 2\sum_{\substack{a\in[m/2]\\b\in[m]\\a+b<m}}\exp\left(-\left(\frac{1}{4}C_{\mathrm{reg},2}^2-2\right)\log m\cdot(m-b)\right).$$

Finally, there are at most m^2 terms in the summation, and $m-b\geq 1$ in all terms, so we finish by bounding, for $C_{\mathrm{reg},2}$ sufficiently large,

$$\leq \lim_{m\to\infty} 2m^2\exp\left(-\left(\frac{1}{4}C_{\mathrm{reg},2}^2-2\right)\log m\right)$$
$$\leq \lim_{m\to\infty} 2\exp\left(-\left(\frac{1}{4}C_{\mathrm{reg},2}^2-4\right)\log m\right) \tag{36}$$

Choosing $C_{\mathrm{reg},2}$ again sufficiently large, this limit will equal zero.

Combining (32) and (36), we have therefore found that, for $C_{\mathrm{reg},2}$ sufficiently large,

$$\lim_{m\to\infty}\sum_{U\subseteq V}\mathbb{P}[A_U^c] = 0.$$

Using this in (31), we have

$$\lim_{m\to\infty}\mathbb{P}[A^c] \leq \lim_{m\to\infty}\mathbb{P}[A_{\deg}^c] + \lim_{m\to\infty}\sum_{U\subseteq V}\mathbb{P}[A_U^c] = 0,$$

giving the result.

5.2 Proof of Proposition 4

In this proof we consider a general ER graph $G \sim \mathcal{G}_{n,p}$ with the minimum node degree denoted by $d_{\min}$ and we show that with high probability G contains as a subgraph, a d_{reg}-regular graph with the same node set, where d_{reg} is asymptotically equal to $d_{\min}$. To this end we make use of the following results concerning the existence of b-factors for general graphs, degree distribution of dense random graphs and node connectivity of dense random graphs. Let $U \subseteq V$; in the following, we denote by $G - U$ the graph obtained from G by removing all the nodes in U together with all the edges incident with nodes in U.

Proposition 7 ([40]). *Let $G = (V, E)$ be a graph and let $f \in \mathbb{Z}^V_{\geq 0}$. For disjoint subsets $S, T \subseteq V$, let $\widetilde{Q}(S, T)$ denote the number of connected components C of $G - (S \cup T)$ such that $f(C) + N(C, T) \equiv 1 \pmod 2$. Then, G has an f-factor if and only if, for all disjoint subsets $S, T \subseteq V$,*

$$f(S) - f(T) + \sum_{v \in T} d(v) - N(S, T) - \widetilde{Q}(S, T) \geq 0.$$

Proposition 8 (Chapter 3 of [21]). *Let $G \sim \mathcal{G}_{n,p}$ for some $p \in (0, 1)$. Denote by $d_{\min}$ the minimum node degree of G. Then, there exists a constant $C > 0$ such that, with high probability, $d_{\min} \geq pn - C\sqrt{n \log n}$.*

Recall that the node connectivity $\kappa(G)$ of a graph G is the minimum number of nodes whose deletion disconnects it, or the total number of nodes for G a complete graph. It is well-known that $\kappa(G) \leq d_{\min}$, where as before $d_{\min}$ denotes the minimum node degree in G. Surprisingly, for ER graphs, the two quantities coincide with high probability.

Proposition 9 (Theorem 1 of [9]). *Let $G \sim \mathcal{G}_{n,p}$ for some $p \in (0, 1)$. Then, with high probability, $\kappa(G) = d_{\min}$.*

We proceed with the proof of Proposition 4. First, for $A \subseteq V$, let $Q(A)$ denote the number of connected components in $G - A$. Clearly, $\widetilde{Q}(S, T) \leq Q(S \cup T)$ for any disjoint $S, T \subseteq V$ (where $\widetilde{Q}(S, T)$ is as defined in the statement of Proposition 7).

We will apply Proposition 7 to the function $f(v) = d_{\text{reg}}$ for all $v \in V$. Without loss of generality, we may suppose d_{reg} is even. In this case, the case $S = T = \emptyset$ is satisfied automatically, so we may suppose $|S| + |T| \geq 1$ in the sequel. Using the above observation, we find that the desired regular subgraph exists provided the following event occurs:

$$A := \bigcap_{\substack{S,T \subseteq V \\ S \cap T = \emptyset \\ |S| + |T| \geq 1}} A_{S,T}, \qquad \text{where } A_{S,T} := \left\{ N(S, T) - \sum_{v \in T} d(v) \leq d_{\text{reg}}(|S| - |T|) - Q(S \cup T) \right\}.$$

We have

$$\sum_{v \in T} d(v) = 2N(T, T) + N(S, T) + N(V \setminus (S \cup T), T),$$

whereby we may rewrite

$$\begin{aligned}
A_{S,T} &= \Big\{ -2N(T, T) - N(V \setminus (S \cup T), T) \leq d_{\text{reg}}(|S| - |T|) - Q(S \cup T) \Big\} \\
&= \Big\{ - \sum_{\{i,j\} \in \binom{T}{2}} (2X_{ij} - \mathbb{E}[2X_{ij}]) - \sum_{\substack{i \in V \setminus (S \cup T) \\ j \in T}} (X_{ij} - \mathbb{E}[X_{ij}]) \\
&\qquad \leq p|T|(|T| - 1) + p|T|(n - |S| - |T|) + d_{\text{reg}}(|S| - |T|) - Q(S \cup T) \Big\} \\
&= \Big\{ - \sum_{\{i,j\} \in \binom{T}{2}} (2X_{ij} - \mathbb{E}[2X_{ij}]) - \sum_{\substack{i \in V \setminus (S \cup T) \\ j \in T}} (X_{ij} - \mathbb{E}[X_{ij}]) \\
&\qquad \leq \Big(p(n - |T|) - C_{\text{reg},1}\sqrt{n \log n} \Big)|S| + (C_{\text{reg},1}\sqrt{n \log n} - p)|T| - Q(S \cup T) \Big\},
\end{aligned}$$

where we have performed the manipulation

$$\begin{aligned}
&p|T|(|T|-1)+p|T|(n-|S|-|T|)+d_{\mathrm{reg}}(|S|-|T|)\\
&\qquad=p|T|(n-|S|-1)+(pn-C_{\mathrm{reg},1}\sqrt{n\log n})(|S|-|T|)\\
&\qquad=p|S|n-p|S|\cdot|T|-p|T|+C_{\mathrm{reg},1}\sqrt{n\log n}(|T|-|S|)\\
&\qquad=p|S|(n-|T|)+C_{\mathrm{reg},1}\sqrt{n\log n}(|T|-|S|)-p|T|\\
&\qquad=\Big(p(n-|T|)-C_{\mathrm{reg},1}\sqrt{n\log n}\Big)|S|+(C_{\mathrm{reg},1}\sqrt{n\log n}-p)|T|.
\end{aligned}$$

Now, we give a lower bound for the right-hand side by considering two cases, depending on the size of $|T|$. If $|T|\le n-2C_{\mathrm{reg},1}p^{-1}\sqrt{n\log n}$, then $p(n-|T|)-C_{\mathrm{reg},1}\sqrt{n\log n}\ge C_{\mathrm{reg},1}\sqrt{n\log n}$, and for sufficiently large n we also have $C_{\mathrm{reg},1}\sqrt{n\log n}-p\ge\frac{1}{4}C_{1,\mathrm{reg}}\sqrt{n\log n}$, so in this case

$$\Big(p(n-|T|)-C_{\mathrm{reg},1}\sqrt{n\log n}\Big)|S|+(C_{\mathrm{reg},1}\sqrt{n\log n}-p)|T|\ge\frac{1}{4}C_{\mathrm{reg},1}\sqrt{n\log n}(|S|+|T|).$$

If, on the other hand, $|T|\ge n-2C_{\mathrm{reg},1}p^{-1}\sqrt{n\log n}$, then, since S and T are disjoint $|S|+|T|\le n$ whereby $|S|\le n-|T|\le 2C_{\mathrm{reg},1}p^{-1}\sqrt{n\log n}$. Therefore, for sufficiently large n, we will have

$$\begin{aligned}
&\Big(p(n-|T|)-C_{\mathrm{reg},1}\sqrt{n\log n}\Big)|S|+(C_{\mathrm{reg},1}\sqrt{n\log n}-p)|T|\\
&\qquad\ge\frac{1}{2}C_{\mathrm{reg},1}n\sqrt{n\log n}-C_{\mathrm{reg},1}\sqrt{n\log n}|S|\ge\frac{1}{2}C_{\mathrm{reg},1}n\sqrt{n\log n}-2C_{\mathrm{reg},1}^2p^{-1}n\log n\\
&\qquad\ge\frac{1}{4}C_{\mathrm{reg},1}n\sqrt{n\log n}\ge\frac{1}{4}C_{\mathrm{reg},1}\sqrt{n\log n}(|S|+|T|).
\end{aligned}$$

Thus the same bound is obtained in either case, and we find that, for sufficiently large n, for all $S,T\subseteq V$ disjoint, we have

$$\begin{aligned}
A_{S,T}\supseteq\Bigg\{&-2\sum_{\{i,j\}\in\binom{T}{2}}(X_{ij}-\mathbb{E}[X_{ij}])-\sum_{\substack{i\in V\setminus(S\cup T)\\ j\in T}}(X_{ij}-\mathbb{E}[X_{ij}])\\
&\le\frac{1}{4}C_{\mathrm{reg},1}\sqrt{n\log n}(|S|+|T|)-Q(S\cup T)\Bigg\}.
\end{aligned}$$

Now, note that we always have $Q(S\cup T)\le n$. Moreover, if $\kappa(G)$ is the node connectivity of G, and $|S|+|T|<\kappa(G)$, then $Q(S\cup T)=1$. And, by Propositions 8 and 9, with high probability $\kappa=d_{\min}\ge pn-C\sqrt{n\log n}$. Let A_{con} denote the event that both the equality $\kappa=d_{\min}$ and the subsequent inequality hold in G. Then, for sufficiently large n, on the event A_{con},

$$Q(S\cup T)\le\left\{\begin{array}{ll}1 & \text{if } |S|+|T|\le pn/2\\ n & \text{otherwise}\end{array}\right\}\le\frac{2}{p}(|S|+|T|),$$

where in the last step we use that S and T are not both empty. Thus, again for sufficiently large n,

$$\begin{aligned}
A_{S,T}\supseteq A_{\mathrm{con}}\cap\Bigg\{&-2\sum_{\{i,j\}\in\binom{T}{2}}(X_{ij}-\mathbb{E}[X_{ij}])-\sum_{\substack{i\in V\setminus(S\cup T)\\ j\in T}}(X_{ij}-\mathbb{E}[X_{ij}])\\
&\le\frac{1}{8}C_{\mathrm{reg},1}\sqrt{n\log n}(|S|+|T|)\Bigg\}.
\end{aligned}$$

Now, returning to the main event A, we note two special cases that make the left-hand side of the definition of the event inside the intersection above equal to zero: (1) if $T=\emptyset$, and (2) if $|T|=1$ and

$S \cup T = V$. Thus we may neglect both of these cases, and write, for sufficiently large n,

$$A \supseteq A_{\mathrm{con}} \cap \bigcap_{\substack{S,T \subseteq V \\ S \cap T = \emptyset \\ |T| \geq 2 \text{ or } |S|+|T| < n}} \Bigg\{ -2 \sum_{\{i,j\} \in \binom{T}{2}} (X_{ij} - \mathbb{E}[X_{ij}]) - \sum_{\substack{i \in V \setminus (S \cup T) \\ j \in T}} (X_{ij} - \mathbb{E}[X_{ij}]) \\ \leq \frac{1}{8} C_{\mathrm{reg},1} \sqrt{n \log n} (|S| + |T|) \Bigg\}.$$

Applying a union bound and Hoeffding's inequality, we may then bound

$$\lim_{n \to \infty} \mathbb{P}[A^c] \leq \lim_{n \to \infty} \mathbb{P}[A_{\mathrm{con}}^c] + \lim_{n \to \infty} \sum_{\substack{S,T \subseteq V \\ S \cap T = \emptyset \\ |T| \geq 2 \text{ or } |S|+|T| < n}} \exp\left(-2 \frac{\frac{1}{64} C_{\mathrm{reg},1}^2 n \log n (|S| + |T|)^2}{2|T|(|T|-1) + |T|(n - |S| - |T|)} \right).$$

The first limit is zero by our previous remark. In analyzing the remaining exponential terms, we first bound the denominator as $2|T|(|T|-1) + |T|(n - |S| - |T|) \leq 3n|T|$, continuing

$$\leq \lim_{n \to \infty} \sum_{\substack{S,T \subseteq V \\ S \cap T = \emptyset \\ |T| \geq 2 \text{ or } |S|+|T| < n}} \exp\left(-\frac{1}{96} C_{\mathrm{reg},1}^2 \frac{\log n \cdot (|S| + |T|)^2}{|T|} \right)$$

and then observe that $(|S| + |T|)^2 / |T| \geq (|S| + |T|)^2 / (|S| + |T|) = |S| + |T|$, whereby

$$\leq \lim_{n \to \infty} \sum_{\substack{S,T \subseteq V \\ S \cap T = \emptyset \\ |T| \geq 2 \text{ or } |S|+|T| < n}} \exp\left(-\frac{1}{96} C_{\mathrm{reg},1}^2 \log n \cdot (|S| + |T|) \right)$$

$$\leq \lim_{n \to \infty} \sum_{\substack{S,T \subseteq V \\ |S|+|T| > 0}} \exp\left(-\frac{1}{96} C_{\mathrm{reg},1}^2 \log n \cdot (|S| + |T|) \right)$$

and introducing scalar variables $a = |S|$ and $b = |T|$ and grouping according to these values, we further bound

$$= \lim_{n \to \infty} \sum_{\substack{a,b \in \{0,\ldots,n\} \\ a+b>0}} \binom{n}{a} \binom{n}{b} \exp\left(-\frac{1}{96} C_{\mathrm{reg},1}^2 \log n \cdot (a+b) \right).$$

To finish, we use that $\binom{n}{a} \leq \exp(a \log n)$ and $\binom{n}{b} \leq \exp(b \log n)$, and there are at most $(n+1)^2$ terms in the outer sum, whereby

$$\leq \lim_{n \to \infty} \sum_{\substack{a,b \in \{0,\ldots,n\} \\ a+b>0}} \exp\left(-\frac{1}{96} C_{\mathrm{reg},1}^2 \log n \cdot (a+b) + \log n \cdot (a+b) \right)$$

$$\leq \lim_{n \to \infty} \exp\left(-\left(\frac{1}{96} C_{\mathrm{reg},1}^2 - 1 \right) \log n + 2 \log(n+1) \right).$$

Setting $C_{\mathrm{reg},1}$ sufficiently large will make the remaining limit equal zero, giving the result.

5.3 Recovery guarantees for the SDP relaxation

In the following we state a set of sufficient conditions under which a well-known SDP relaxation of the min-bisection problem recovers the planted bisection under the SBM. This result follows immediately from Lemma 3.13 of [5] (see [5] for further details about the SDP relaxation). To prove this result we make use of Bernstein's inequality (see Theorem 2.8.4 in [41]).

Proposition 10. *Suppose $p = p(n) \in (0,1)$, $q = q(n) \in (0,1)$ are such that the following conditions hold:*

1. $\frac{\log n}{3n} < p < \frac{1}{2}$ *for sufficiently large n.*
2. *There exists $\epsilon > 0$ such that $p - q \geq (12+\epsilon)\sqrt{p \cdot \frac{\log n}{n}}$ for sufficiently large n.*

Then with high probability, the SDP relaxation recovers the planted bisection in a graph drawn from $\mathcal{G}_{n,p,q}$.

Proof. Define

$$\deg_{\text{in}}(v) = \begin{cases} N(\{v\}, V_1) & \text{if } v \in V_1 \\ N(\{v\}, V_2) & \text{if } v \in V_2 \end{cases}, \qquad \deg_{\text{out}}(v) = \begin{cases} N(\{v\}, V_2) & \text{if } v \in V_1 \\ N(\{v\}, V_1) & \text{if } v \in V_2 \end{cases}.$$

Then, by Lemma 3.13 of [5], it suffices to show that for any constant $\Delta > 0$, with high probability

$$\min_v \{\deg_{\text{in}}(v) - \deg_{\text{out}}(v)\} \geq \frac{\Delta}{\sqrt{\log n}} \mathbb{E}\left[\deg_{\text{in}}(v_0) - \deg_{\text{out}}(v_0)\right]$$

for an arbitrary fixed node v_0 (the expectation on the right-hand side does not depend on this choice). Noting that

$$\mathbb{E}\left[\deg_{\text{in}}(v_0) - \deg_{\text{out}}(v_0)\right] = \frac{p-q}{2}n - p,$$

it suffices to show the weaker statement that, again for any constant $\Delta > 0$, with high probability

$$\min_v \deg_{\text{in}}(v) - \max_v \deg_{\text{out}}(v) \geq \Delta \frac{p-q}{2} \frac{n}{\sqrt{\log n}}.$$

Note that for any v, $\deg_{\text{in}}(v)$ is a sum of $\frac{n}{2} - 1$ random Bernoulli variables with mean p, and therefore with variance $p(1-p) \leq p$. Therefore, using Bernstein's inequality,

$$\begin{aligned} \mathbb{P}\left[\deg_{\text{in}}(v) - \mathbb{E}[\deg_{\text{in}}(v)] \leq -t\sqrt{pn\log n}\right] &\leq \exp\left(-\frac{\frac{1}{2}t^2 pn \log n}{\frac{1}{2}pn + \frac{2}{3}\sqrt{pn\log n}}\right) \\ &\leq \exp\left(-\frac{t^2\sqrt{pn}\log n}{\sqrt{pn} + \frac{4}{3}\sqrt{\log n}}\right) \\ &\leq \exp\left(-\frac{1}{5}t^2 \log n\right). \end{aligned}$$

Thus since $\mathbb{E}[\deg_{\text{in}}(v)] = p(\frac{n}{2} - 1)$, taking $t = 3$, and using a union bound over v, it follows that, with high probability,

$$\min_v \deg_{\text{in}}(v) \geq \frac{1}{2}pn - 3\sqrt{pn\log n}.$$

By a symmetric argument for $d_{\text{out}}(v)$, with high probability,

$$\max_v \deg_{\text{out}}(v) \leq \frac{1}{2}qn + 3\sqrt{pn\log n}.$$

Letting $t = \max\{t_{\text{in}}, t_{\text{out}}\}$, we find that, with high probability,

$$\min_v \deg_{\text{in}}(v) - \max_v \deg_{\text{out}}(v) \geq \frac{p-q}{2}n - 6\sqrt{pn\log n}.$$

Thus it suffices to show that under the assumptions in the statement, for any $\Delta > 0$ and sufficiently large n,

$$\frac{p-q}{2}n - 6\sqrt{pn\log n} \geq \frac{p-q}{2}n \cdot \frac{\Delta}{\sqrt{\log n}}.$$

This follows since, by Assumption (2), there exists $\delta > 0$ such that $6\sqrt{pn\log n} \leq (1-\delta) \cdot \frac{1}{2}(p-q)n$ for sufficiently large n. □

Proposition 10 in particular implies that, if $p = \alpha n^{-\omega}$ and $q = \beta n^{-\omega}$ for any $\omega \in (0,1)$, the SDP relaxation recovers the planted bisection with high probability provided that $p > q$ (or equivalently $\alpha > \beta$). When $p = \alpha\frac{\log n}{n}$ and $q = \beta\frac{\log n}{n}$, by Lemma 4.11 of [5], recovery is guaranteed with high probability provided that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$.

5.4 Distance distributions in random graphs

Consider a graph graph $G \sim \mathcal{G}_{n,p,q}$. In this section, we present lower and upper bounds on the average distance $\rho_{\text{avg}}(G)$ and the diameter $\rho_{\max}(G)$ of G in very dense, dense, and logarithmic regimes. We use these results to prove our non-recovery conditions given by Theorems 7 and 8. The distance distributions of ER graphs have been studied extensively in the literature [8, 12, 13, 38]. To obtain similar results for the SBM, we use the following tool to relate the distance distributions of SBM graphs to those of ER graphs

Proposition 11. *Let $0 < q \le p < 1$ and let n be an even number. Then, there exists a probability distribution over triples of graphs (G_1, G_2, G_3) on a mutual set V of n nodes such that the following conditions hold:*

1. *The marginal distributions of G_1, G_2, and G_3 are $\mathcal{G}_{n,q}$, $\mathcal{G}_{n,q,p}$, and $\mathcal{G}_{n,p}$, respectively.*
2. *With probability 1, G_1 is a subgraph of G_2 and G_2 is a subgraph of G_3. Consequently, for all $i, j \in V$, $\rho_{G_1}(i,j) \ge \rho_{G_2}(i,j) \ge \rho_{G_3}(i,j)$.*

Proof. We describe a procedure for sampling the three graphs together. Let V be their mutual node set, and fix a uniformly random bisection $V = V_1 \sqcup V_2$ with $|V_1| = |V_2| = \frac{n}{2}$.

The key observation is the following. Let $\text{Ber}(a)$ denote the Bernoulli distribution with probability a. Then, if $X \sim \text{Ber}(a)$ and $Y \sim \text{Ber}(b)$ are sampled independently, then the binary OR of X with Y has the law $\text{Ber}(a + b - ab)$. Now, define $a = \frac{p-q}{1-q}$. Since $q + \frac{p-q}{1-q} - q \cdot \frac{p-q}{1-q} = p$, we have that if $X \sim \text{Ber}(q)$ and $Y \sim \text{Ber}(a)$, then the binary OR of X with Y has law $\text{Ber}(p)$.

We now describe a procedure for sampling the desired triple of graphs. First, sample $G_1 \sim \mathcal{G}_{n,q}$. Next, for each $i \in V_1$ and $j \in V_2$, sample $X_{ij} \sim \text{Ber}(a)$, independently of G_1. Let G_2 have all of the edges of G_1, and also an edge between i and j if $X_{ij} = 1$. By the previous observation, the law of G_2 is $\mathcal{G}_{n,q,p}$.

Similarly, for each $i, j \in V_1$ and $i, j \in V_2$, sample $Y_{ij} \sim \text{Ber}(a)$, independently of G_1 and G_2. Let G_3 have all of the edges of G_2, and also an edge between i and j if $Y_{ij} = 1$. By the previous observation, the law of G_3 is $\mathcal{G}_{n,p}$. □

We now briefly review the existing results on the distance distributions of ER graphs. It will be convenient for us to give these results in slightly different form than in the original references, so we derive the statements we will need from prior work below. Below, for $x > 0$, we write $\lfloor x \rfloor$ for the "floor function" or greatest integer not exceeding x, and $\{x\} := x - \lfloor x \rfloor$ for the "fractional part" of x.

Proposition 12. *Suppose $p = \alpha n^{-\omega}$ for some $\omega \in [0, 1)$ and $\alpha > 0$, with $\alpha < 1$ if $\omega = 0$. Let $G \sim \mathcal{G}_{n,p}$. Then,*

$$\lim_{n\to\infty} \mathbb{P}\left[\rho_{\max}(G) = 1 + \left\lfloor \frac{1}{1-\omega} \right\rfloor\right] = 1.$$

Proof. First, suppose $\omega \in [0, \frac{1}{2})$. Then, our task is to show that $\rho_{\max} = 2$ with high probability. By Corollary 10.11(i) of [8], it suffices to check that $p^2 n - 2\log n \to \infty$ and $n^2(1-p) \to \infty$. The latter clearly holds for any $\omega > 0$, and holds for $\omega = 0$ since in that case $p = \alpha \in (0, 1)$. For the former, we have $p^2 n = \alpha n^{1-2\omega}$ with $\alpha > 0$ and $\omega < \frac{1}{2}$, and the result follows.

Next, suppose $\omega \in [\frac{1}{2}, 1)$. Define $d := 1 + \lfloor \frac{1}{1-\omega} \rfloor \ge 3$. By Corollary 10.12(i) of [8], it suffices to check that $d^{-1}\log n - 3\log\log n \to \infty$, $p^d n^{d-1} - 2\log n \to \infty$, and $p^{d-1} n^{d-2} - 2\log n \to -\infty$. The first condition follows since d is a constant. For the second condition, we calculate

$$p^d n^{d-1} = \alpha^d n^{-\omega(1+\lfloor \frac{1}{1-\omega} \rfloor) + \lfloor \frac{1}{1-\omega} \rfloor}.$$

Manipulating the exponent, we have

$$\begin{aligned} -\omega(1 + \lfloor \frac{1}{1-\omega} \rfloor) + \lfloor \frac{1}{1-\omega} \rfloor &= -\omega + (1-\omega)\lfloor \frac{1}{1-\omega} \rfloor \\ &= (1-\omega)\left(1 - \left\{\frac{1}{1-\omega}\right\}\right). \end{aligned}$$

We have $(1-\omega) \in (0,1)$ and $1 - \{\frac{1}{1-\omega}\} \in (0,1]$, so $p^d n^{d-1} = \alpha^d n^\delta$ for some $\delta > 0$, so the second condition holds. Finally, for the third condition we have

$$p^{d-1} n^{d-2} = \alpha^{d-1} n^{-\omega(1+\lfloor \frac{1}{1-\omega} \rfloor) + \lfloor \frac{1}{1-\omega} \rfloor - (1-\omega)} = \alpha^{d-1} n^{-(1-\omega)\{\frac{1}{1-\omega}\}} = \alpha^{d-1} n^{-\delta'}$$

for some $\delta' \geq 0$. Thus the third condition also holds, and the result follows. □

Proposition 13. *Suppose $p = \alpha \frac{\log n}{n}$ for some $\alpha > 1$. Let $G \sim \mathcal{G}_{n,p}$. Then, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left[\rho_{\max}(G) \in \left((1-\epsilon)\frac{\log n}{\log\log n}, (1+\epsilon)\frac{\log n}{\log\log n}\right)\right] = 1.$$

Proof. We note that the asymptotic $\frac{\log n}{\log pn} \sim \frac{\log n}{\log\log n}$ holds as $n \to \infty$, since $\log pn = \log\alpha + \log\log n$. Thus it suffices to show that

$$\lim_{n\to\infty} \mathbb{P}\left[\rho_{\max}(G) \in \left((1-\epsilon)\frac{\log n}{\log pn}, (1+\epsilon)\frac{\log n}{\log pn}\right)\right] = 1.$$

We explicitly make this trivial rewriting to put our result in the form usually used in the literature. The result in this form then follows directly from Theorem 4 of [12]. □

Proposition 14 (Theorem 1 of [38])**.** *Suppose $p = \alpha n^{-\omega}$ for some $\omega \in (0,1)$ and $\alpha > 0$. Define the quantity*

$$\mu = \mu(\omega, \alpha) = \left\lceil \frac{1}{1-\omega} \right\rceil + \mathbf{1}\left\{\frac{1}{1-\omega} \in \mathbb{N}\right\} \exp\left(-\alpha^{\frac{1}{1-\omega}}\right).$$

Let $G \sim \mathcal{G}_{n,p}$. Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\rho_{\text{avg}}(G) \in ((1-\epsilon)\mu, (1+\epsilon)\mu)\right] = 1.$$

Proposition 15 (Theorem 1 of [13])**.** *Suppose $p = \alpha \frac{\log n}{n}$ for some $\alpha > 1$. Let $G \sim \mathcal{G}_{n,p}$. Then, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left[\rho_{\text{avg}}(G) \in \left((1-\epsilon)\frac{\log n}{\log pn}, (1+\epsilon)\frac{\log n}{\log pn}\right)\right] = 1.$$

Utilizing the above results on distance distribution of ER graphs together with Proposition 11, in the following we obtain similar results for the SBM.

Proposition 16. *Suppose $0 < q < p < 1$ are constants. Let $G \sim \mathcal{G}_{n,p,q}$. Then, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left[(1-\epsilon)\left(2 - \frac{p+q}{2}\right) \leq \rho_{\text{avg}}(G) \leq \rho_{\max}(G) = 2\right] = 1.$$

Proof. By Proposition 11 and Proposition 12, with high probability $\rho_{\max}(G) = 2$ when $G \sim \mathcal{G}_{n,p,q}$. On this event, every adjacent pair of nodes in G has distance 1, and every non-adjacent pair of nodes has distance 2. Therefore, we may explicitly calculate the average distance: when $\rho_{\max}(G) = 2$, then

$$\rho_{\text{avg}}(G) = \frac{2}{n^2}\left(1 \cdot |E| + 2 \cdot \left(\binom{n}{2} - |E|\right)\right) = 2\left(1 - \frac{1}{n}\right) - \frac{2}{n^2}|E|.$$

By Hoeffding's inequality, for any fixed $\delta > 0$, with high probability $|E| \leq (1+\delta) \cdot \frac{p+q}{2}\binom{n}{2}$ (where the second factor on the right-hand side is $\mathbb{E}|E|$). The result then follows after substituting and choosing δ sufficiently small depending on the given ϵ. □

Proposition 17. *Suppose $p = \alpha n^{-\omega}$ and $q = \beta n^{-\omega}$ for $\alpha, \beta > 0$ and $\omega \in (0,1)$ such that $\frac{1}{1-\omega} \notin \mathbb{N}$. Let $G \sim \mathcal{G}_{n,p,q}$. Then, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left[(1-\epsilon)\left\lceil \frac{1}{1-\omega} \right\rceil \leq \rho_{\text{avg}}(G) \leq \rho_{\max}(G) = \left\lceil \frac{1}{1-\omega} \right\rceil\right] = 1.$$

Proof. The result follows from using Proposition 11 to compare to ER graphs, and then applying Proposition 14 and Proposition 12. □

Proposition 18. *Suppose* $p = \alpha n^{-\omega}$ *and* $q = \beta n^{-\omega}$ *for* $\alpha > \beta > 0$ *and* $\omega \in (0,1)$ *such that* $\frac{1}{1-\omega} \in \mathbb{N}$. *Let* $G \sim \mathcal{G}_{n,p,q}$. *Then, for any* $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[(1-\epsilon)\left(\frac{1}{1-\omega} + \exp\left(-\alpha^{\frac{1}{1-\omega}}\right)\right) \leq \rho_{\text{avg}}(G) \leq \rho_{\max}(G) = \frac{1}{1-\omega} + 1\right] = 1.$$

Proof. The result follows from using Proposition 11 to compare to ER graphs, and then applying Proposition 14 and Proposition 12. □

Proposition 19. *Suppose* $p = \alpha\frac{\log n}{n}$ *and* $q = \beta\frac{\log n}{n}$ *for* $\alpha > \beta > 0$ *satisfying* $\frac{\alpha+\beta}{2} > 1$. *Let* $G \sim \mathcal{G}_{n,p,q}$. *Then, for any* $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[(1-\epsilon)\frac{\log n}{\log\log n} \leq \rho_{\text{avg}}(G) \leq \rho_{\max}(G) \leq (1+\epsilon)\frac{\log n}{\log\log n}\right] = 1.$$

Proof. The middle inequality holds for any graph G. For the left inequality, we use Proposition 11 to construct a graph $G' \sim \mathcal{G}_{n,p}$ coupled to G such that G is a subgraph of G'. Then, $\rho_{\text{avg}}(G) \geq \rho_{\text{avg}}(G')$, and the result then follows by Proposition 15. The right inequality follows from a standard result on the diameter of the SBM, which may be obtained by repeating the analysis of the size of the neighborhood of a node given in Section 4.5.1 of [1] in the case of logarithmic average degree rather than constant average degree. □

Acknowledgements The authors would like to thank Afonso Bandeira for fruitful discussions about community detection and the recovery properties of SDP relaxations.

References

[1] E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18:1 – 86, 2018.

[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62:471 – 487, 2016.

[3] E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48:1452–1474, 2020.

[4] D. Avis and J. Umemoto. Stronger linear programming relaxations of max-cut. *Mathematical Programming*, 97(3):451–469, 2003.

[5] A. S. Bandeira. Random Laplacian matrices and convex relaxations. *Foundations of Computational Mathematics*, 18(2):345–379, 2018.

[6] P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[7] M. Boedihardjo, S. Deng, and T. Strohmer. A performance guarantee for spectral clustering. *SIAM Journal on Mathematics of Data Science*, 3(1):369–387, 2021.

[8] B. Bollobás. *Random graphs*. Number 73 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, second edition, 2001.

[9] B. Bollobás and A. Thomason. Random graphs of small order. *Annals of Discrete Mathematics*, 28:47–97, 1985.

[10] T. Carson and R. Impagliazzo. Hill-climbing finds random planted bisections. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 903–909. Society for Industrial and Applied Mathematics, 2001.

[11] M. Charikar, K. Makarychev, and Y. Makarychev. Integrality gaps for Sherali-Adams relaxations. In *STOC '09: Proceedings of the forty-first annual ACM Symposium on Theory of Computing*, pages 283–292, 2009.

[12] F. Chung and L. Lu. The diameter of sparse random graphs. *Advances in Applied Mathematics*, 26(4):257–279, 2001.

[13] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[14] W. F. De la Vega and C. Kenyon-Mathieu. Linear programming relaxations of maxcut. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 53–61, 2007.

[15] A. De Rosa and A. Khajavirad. Efficient joint object matching via linear programming. *arXiv:2108.11911*, 2021.

[16] A. De Rosa and A. Khajavirad. The ratio-cut polytope and K-means clustering. *SIAM Journal on Optimization*, 32:173–203, 2022.

[17] A. Del Pia and A. Khajavirad. Rank-one Boolean tensor factorization and the multilinear polytope. *arXiv:2202.07053*, 2022.

[18] A. Del Pia and M. Ma. k-median: exact recovery in the extended stochastic ball model. *Preprint, arXiv:2109.02547*, 2021.

[19] M. M. Deza and M. Laurent. *Geometry of cuts and metrics*, volume 15 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin Heidelberg New York, 1997.

[20] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[21] A. Frieze and M. Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[22] M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete problems. In *Proceedings of the sixth annual ACM symposium on Theory of computing*, pages 47–63, 1974.

[23] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788 – 2797, 2016.

[24] Samuel B Hopkins, Tselil Schramm, and Luca Trevisan. Subexponential lps approximate max-cut. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–953. IEEE, 2020.

[25] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. Probably certifiably correct k-means clustering. *Mathematical Programming*, 165:605–642, 2017.

[26] R. Krauthgamer and U. Feige. A polylogarithmic approximation of the minimum bisection. *SIAM Review*, 48(1):99–130, 2006.

[27] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. *In Foundations of Computer Science (FOCS)*, pages 299–308, 2010.

[28] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46:787–832, 1999.

[29] X. Li, Y. Li, S. Ling, T. Strohmer, and K. Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *Mathematical Programming*, 179:295–341, 2020.

[30] S. Ling and T. Strohmer. Certifying global optimality of graph cuts via semidefinite relaxation: A performance guarantee for spectral clustering. *Foundations of Computational Mathematics*, 2019.

[31] O. L. Mangasarian. Uniqueness of solution in linear programming. *Linear Algebra and its Applications*, 25:151–162, 1979.

[32] D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

[33] A. Moitra, W. Perry, and A. S. Wein. How robust are reconstruction thresholds for community detection? In *STOC '16: Proceedings of the forty-eighth annual ACM Symposium on Theory of Computing*, pages 828–841, 2016.

[34] R. O'Donnell and T. Schramm. Sherali–adams strikes back. *Theory of Computing*, 17(1):1–30, 2021.

[35] S. Poljak and Z. Tuza. The expected relative error of the polyhedral approximation of the max-cut problem. *Operations Research Letters*, 16:191 – 198, 1994.

[36] F. Ricci-Tersenghi, A. Javanmard, and A. Montanari. Performance of a community detection algorithm based on semidefinite programming. *Journal of Physics: Conference Series*, 699:12015–12025, 2016.

[37] A. Schrijver. *Combinatorial Optimization. Polyhedra and Efficiency.* Springer-Verlag, Berlin, 2003.

[38] N. Shimizu. The average distance of dense homogeneous random graphs. Mathematical engineering technical report, Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Nov 2017.

[39] T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[40] W. T. Tutte. The factors of graphs. *Canadian Journal of Mathematics*, 4:314–328, 1952.

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