Proceedings of Machine Learning Research vol 291:1–29, 2025 38th Annual Conference on Learning Theory

Robust Algorithms for Recovering Planted r-Colorable Graphs

Anand Louis ANANDL@IISC.AC.IN
Indian Institute of Science, Bangalore

Rameesh Paul RAMEESHPAUL@IISC.AC.IN
Indian Institute of Science, Bangalore

Prasad Raghavendra RAGHAVENDRA@BERKELEY.EDU
University of California, Berkeley

Editors: Nika Haghtalab and Ankur Moitra

Abstract

The planted clique problem is a fundamental problem in study of algorithms and has been extensively studied in various random and semirandom models. It is known that a clique planted in a random graph can be efficiently recovered if the size of the clique is above the conjectured computational threshold of $\Omega_p\left(\sqrt{n}\right)$. A natural question that arises then is what other planted structures can be efficiently recovered?

In this work, we investigate this question by considering random planted and semirandom models for the r-coloring problem. We study the following model of instances,

- Choose $S \subseteq V$ (of size k) and plant an arbitrary r-colorable graph in the subgraph induced on S.
- For each pair of vertices in $((V \setminus S) \times (V \setminus S))$, add an edge independently with probability p.
- Allow an adversary to add an arbitrary set of edges between vertices in S and vertices in $V \setminus S$.

Our main result is a polynomial-time algorithm that recovers most of the vertices of the planted r-colorable graph when $k \geq cr\sqrt{n/p}$, for some constant c. Our key technical contribution is a novel SDP relaxation and a rounding algorithm for this problem. Our algorithm is also robust to presence of a monotone adversary that can insert edges in the graph induced on $V \setminus S$.

Keywords: Planted Graphs, Graph Coloring, Semirandom Models, SDP Relaxations, Lovász-Theta Function, Strong Adversary, Monotone Adversary, Robust Recovery

1. Introduction

The *planted clique/independent set* problem, first studied by Jerrum (1992) is a central problem in average-case algorithm design. In this problem, a clique/independent set S of size k is planted in an otherwise oblivious Erdős-Rényi random graph denoted $G(n, p)$. The planted structure can be viewed as a signal masked by the background noise of the random $G(n, p)$ graph, where the parameter k represents the strength of this signal. For regimes where $k \geq (2 + \varepsilon) \log_{1/p} n$ (for an arbitrarily small constant $\varepsilon > 0$), the planted clique emerges as the unique k-clique within the graph with high probability. Consequently, in such regimes, the task of identifying the largest clique in the planted graph becomes equivalent to recovering the planted clique itself.

Despite its simplicity, the planted clique problem exhibits a large statistical-computational gap. While information-theoretically, the planted clique is uniquely identifiable for $k \geq (2 + \varepsilon) \log_{1/p} n$ for any constant $\varepsilon > 0$, all known polynomial-time algorithms for the problem computationally succeed only when $k = \Omega_p\left(\sqrt{n}\right)$. There are numerous works (e.g., Alon et al. (1998); Feige and

Krauthgamer (2000); Feige and Kilian (2001); Feige and Ron (2010); Dekel et al. (2014); Deshpande and Montanari (2015)) that design recovery algorithms in this regime using techniques ranging from greedy heuristics to spectral methods, convex relaxations and message-passing algorithms. Improving upon this conjectured computational threshold stands as a fundamental open problem in this area.

On the other hand, lower bounds for restricted class of algorithms, such as statistical query algorithms (Feldman et al. (2017)) and Sum-of-Squares hierarchy (Barak et al. (2019)) suggest that breaking the $\sqrt{n}$ threshold may be impossible. In fact, the *Planted Clique Hypothesis* conjectures that even the "seemingly easier" task of distinguishing a random $G(n, p)$ graph from a graph with a planted clique of size $k = o(\sqrt{n})$ is computationally hard.

This motivates the broader question: *What other kinds of planted subgraphs can be detected and/or recovered?* In a recent work, Kothari et al. (2023) investigate the detection, recovery, and refutation tasks for planted r-colorable graphs in $G(n, 1/2)$ graphs. They present a polynomial time algorithm that achieves recovery when $k = \Omega\left(r\sqrt{n \log n}\right)$. However, their algorithm critically relies on the assumption that the planted graph is a *random* r-colorable graph. They also show that under the Planted Clique Hypothesis, no efficient algorithm can recover when $k = o\left(r\sqrt{n}\right)$.

The problem of recovering planted r-colorable graphs can also be viewed through the lens of average case algorithm design for the partial r-coloring problem, a robust variant of graph coloring problem where a large fraction of graph is r-colorable (see Section 2.3 for details). Studying random and semirandom models may offer valuable insight into the structure of hard instances and guide the design of new algorithms for the problem. This aligns with the broader agenda of *Beyond Worst-Case Algorithm Design*, and this approach has found success for other graph optimization problems, (see e.g., the works of Bhaskara et al. (2010); Błasiok et al. (2024)). Given this motivation, it is important to consider if we can relax the strong assumptions about the planted r-colorable graph in the result of Kothari et al. (2023), specifically the assumptions of random edges between vertices in different color classes, random edges in the graph induced on $S \times (V \setminus S)$ and $V \setminus S$, and equal-sized color classes, and still hope for efficient recovery?

There are numerous other works that study random planted models for a host of problems such as planted clusters aka Stochastic Block Model problem (Abbe and Sandon (2015); Abbe et al. (2016); Chen and Xu (2016); Hajek et al. (2016a)), planted dense subgraphs (Bhaskara et al. (2010); Hajek et al. (2016c,d)), planted graph partitioning (McSherry (2001); Makarychev et al. (2012, 2014)), planted sparse vertex cuts (Louis and Venkat (2018)), planted bipartite graphs (Kumar et al. (2022); Rotenberg et al. (2024)), planted dense cycles (Mao et al. (2023)) among others. Among these, the most relevant to the planted r-colorable subgraph problem are the works on planted dense subgraphs and the works on planted bipartite graphs (they correspond to $r = 2$ case). However, recovery guarantees in these works typically rely on the strong structural assumptions. For example, the recovery algorithms in Hajek et al. (2016b,c,d) for the planted dense subgraphs depend crucially on the planted subgraph having edge density significantly different from the random graph. Similarly, for recovering planted bipartite graphs, the work Rotenberg et al. (2024) assumes a density gap between planted and random subgraphs. The work Kumar et al. (2022) does not assume any such gap but requires that the planted bipartite subgraph is regular. Moreover, both Kumar et al. (2022) and Kothari et al. (2023) require the signal strength k to be at least $\Omega_p\left(r\sqrt{n \log n}\right)$ for efficient recovery, whereas for the planted clique problem, $k = \Omega\left(\sqrt{n}\right)$ suffices. We also note that in terms of robustness, Kothari et al. (2023) doesn't consider an adversary while Kumar et al. (2022) allows for a monotone adversary that can insert edges in the graph induced on $V \setminus S$ and $S \times (V \setminus S)$.

Consequently, a pertinent open question arises: *Can we efficiently recover an arbitrary planted r-colorable graph for $k = \Omega_p\left(r\sqrt{n}\right)$, without imposing any additional assumptions on the planted subgraph, while potentially allowing stronger adversaries, and even permitting sparse regimes where $p = o(1)$?*.

In this work, we affirmatively answer the question, thereby matching the lower bound established in Kothari et al. (2023). Our main result is a polynomial time algorithm that recovers most of the vertices of an arbitrary planted r-colorable subgraph when the size of the planted set S is at least $\Omega_p\left(r\sqrt{n}\right)$ (see Theorem 3). Crucially, our algorithm doesn't make any additional assumptions on the planted r-colorable subgraph and also works in the sparse $p = o(1)$ regimes. Notably, our algorithm is also robust to a strong adversary that can examine the random graph induced on $V \setminus S$ and then add arbitrary edges to the graph induced on $S \times (V \setminus S)$. Further, we can also tolerate a monotone adversary that may insert additional edges in the graph induced on $V \setminus S$. Our key technical contribution is a novel semidefinite programming (SDP) formulation and a corresponding rounding algorithm for the problem.

1.1. Models and Notation

In a similar spirit to the planted clique problem, we define our *random planted model* for planting a r-colorable graph in an oblivious random graph.

Definition 1 (Random Planted Model) *For a given set of parameters n, k, p, and r, we construct an instance of the planted r-colorable subgraph in a random graph as follows:*

1. *Let V denote a set of n vertices. Choose an arbitrary subset $S \subset V$ such that $|S| = k$.*
2. *Add edges within S in an arbitrary manner to form a r-colorable subgraph. Denote the resulting color classes as $S_1, S_2, \ldots, S_r$, where $|S_1| = k_1, |S_2| = k_2, \ldots, |S_r| = k_r$, and $k = k_1 + k_2 + \cdots + k_r$. Let d represent the average degree of G_S, the graph induced on S.*
3. *For each pair of vertices in $S \times (V \setminus S)$, independently add an edge at random with probability p. Similarly, for each pair of vertices in $(V \setminus S) \times (V \setminus S)$, independently add an edge at random with probability p.*

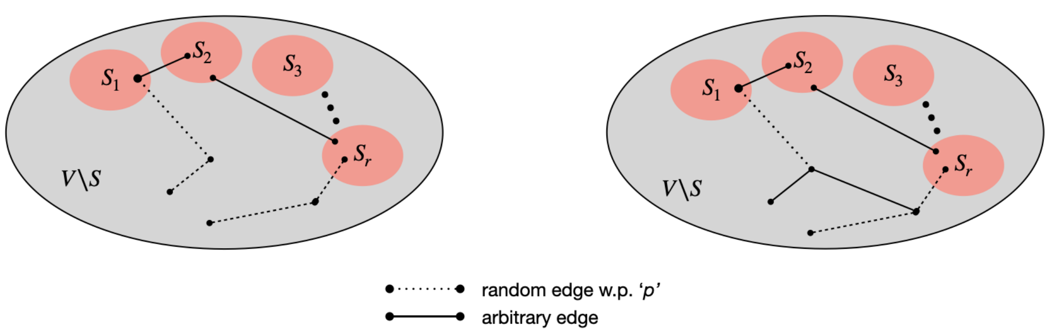

Figure 1: Random Planted Model (left) and Semirandom Model (right)

Next, we consider a semirandom model for planting a r-colorable subgraph in presence of a monotone aka "helpful" adversary (along lines of Blum and Spencer (1995); Feige and Krauthgamer (2000)); and a "strong" adversary.

Definition 2 (Semirandom Model) *Given n, k, r, p, an instance of the semirandom r-colorable graph is constructed as follows,*

1. *Let V denote a set of n vertices. Choose an arbitrary subset $S \subset V$ such that $|S| = k$.*
2. *Add edges within S in an arbitrary manner to form a r-colorable subgraph. Denote the resulting color classes as $S_1, S_2, \ldots, S_r$, where $|S_1| = k_1, |S_2| = k_2, \ldots, |S_r| = k_r$, and $k = k_1 + k_2 + \cdots + k_r$. Let d represent the average degree of G_S, the graph induced on S.*
3. *For each pair of vertices in $(V \setminus S) \times (V \setminus S)$, add edge independently with probability p.*
4. *Allow an adversary to add an arbitrary subgraph between the pair of vertices in $S \times (V \setminus S)$.*
5. *Allow an adversary to arbitrarily insert additional edges among $(V \setminus S) \times (V \setminus S)$.*

Notation. We use $G(n, p)$ to denote a random graph, where an edge is independently added between every pair of vertices i.e. $\{i, j\} \in E$ with probability p. Given a graph G and a set $S \subseteq V$, let G_S denote the subgraph induced on S.

For a matrix $M \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, we define the quadratic form of $\mathbf{x}$ with respect to the matrix M as $\mathbf{x}^T M \mathbf{x}$. We define the Rayleigh quotient of a vector $\mathbf{x}$ with respect to a matrix M denoted $R_M(\mathbf{x})$ as $R_M(\mathbf{x}) = \frac{\mathbf{x}^T M \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$. We use with high probability (also abbreviated as w.h.p.) to indicate that the probability is at least $1 - 1/\mathsf{poly}(n)$.

1.2. Our Results

Benchmarks. We now present our results for both the random planted model and semirandom model. However, to fully appreciate the results, it is crucial to identify our goalpost: when can we hope to recover the planted r-colorable subgraph, and what are inherent limitations we may encounter?

An r-colorable graph can be viewed as a union of r independent sets, which makes this problem a generalization of the planted clique/independent set problem. Accordingly, we do not anticipate recovery to be possible in the regimes of $k = o(\sqrt{n})$ (Feldman et al. (2017); Barak et al. (2019)). Additionally, as discussed earlier, the work Kothari et al. (2023) also rules out recovery (assuming the Planted Clique Hypothesis) for regimes of $k = o\left(r\sqrt{n}\right)$.

It is well-known (see Kucera (1995)) that when k becomes as large as $\Omega(\sqrt{n \log n})$ we can recover the planted clique (w.h.p.) by a simple greedy algorithm that picks out the vertices with large degrees. Kothari et al. (2023) extends this result to planted r-colorable graphs, showing that recovery is possible for $k = \Omega(r\sqrt{n \log n})$. However this extension crucially assumes that the edges between color classes are random. In contrast, our models in Definition 1 and Definition 2 allow arbitrary edges between color classes, and hence the greedy algorithm no longer works (see "Greedy Algorithms" paragraph in Section 2.2 for more discussion). In the context of techniques, we note that for the planted clique/independent set problem, there exists spectral algorithms (Alon et al. (1998)) and SDP based robust algorithms (Feige and Krauthgamer (2000); Feige and Kilian (2001)) that recover the planted clique/independent set S in the regimes where $k = \Omega_p(\sqrt{n})$.

While the planted clique problem is typically studied for $p = 1/2$, it is well-understood for other regimes of p as well (see Feige and Grinberg (2020) for details). Similarly, in this work we will also allow for sparse planted graphs (regimes of $p = o(1)$).

Recovering Planted r-Colorable Graphs. The key takeaway from our work is that we can (approximately) recover a planted r-colorable graph for a wide range of parameters of interest ($k = \Omega_p(r\sqrt{n})$, also allowing $p = o(1)$). Furthermore, our algorithm is also robust to presence of a "strong" adversary and a monotone adversary, as described in Definition 2. Our main result is an SDP based algorithm with the following guarantees,

Theorem 3 (Informal Version of Theorem 9 and Corollary 10) *Given a graph $G = (V, E)$ generated according to Definition 2 and fix any $\delta > 0$. Then for $k \geq c_\delta r\sqrt{n/p}$) (where c_δ is a function of δ), there is a polynomial-time algorithm that outputs a set T such that with high probability (over randomness of the input) it follows that,*

- $|T \cap S| \geq (1 - \delta)k$, *and*
- $|T| \leq (1 + \delta)k$.

We note that our result matches the lower bound established in Kothari et al. (2023).

Remark 4 *The lower bound of $k = o(r\sqrt{n})$, under the Planted Clique Hypothesis, also applies to our models defined in Definition 1 and Definition 2. This is because our models generalize the Kothari et al. (2023) model by removing assumptions on the planted subgraph which can only make the recovery problem harder.*

We discuss the results of Kothari et al. (2023) in detail in "Lower Bounds" paragraph in Section 2.3

Remark 5 *The semirandom model defined in Definition 2 strictly generalizes the random planted model in Definition 1 Consequently, our results in Theorem 3 are also applicable for graphs generated by Definition 1.*

Below, we compare our results with other closely related works such as Kumar et al. (2022); Kothari et al. (2023),

Metrics	**Comparisions with other closely related works**		
	Kumar et al. (2022)	**Kothari et al. (2023)**	**Our Results**
Assumptions on S	$r = 2$ (bipartite graphs) S is d-regular	edges between color classes are random. $\|S_1\|, \|S_2\|, \ldots \|S_r\| = k/r$	no assumptions
Guarantees	recovers planted S for $k = \Omega_p(\sqrt{n \log n})$	recovers planted S for $k = \Omega_p(r\sqrt{n \log n})$	recovers $(1 - \delta)$ fraction of planted S for $k = \Omega_{p,\delta}(r\sqrt{n})$
Robustness	monotone adversary on $S \times (V \setminus S)$ and $V \setminus S$	no adversary considered	arbitrary adversary on $S \times (V \setminus S)$ and monotone adversary on $V \setminus S$

Remark 6 *For graphs generated by Definition 1, and model considered by Kumar et al. (2022), we can use the "clean up" ideas from Kumar et al. (2022) (refer Corollary 19 for details) to also fully recover the planted set for $k = \Omega_p(\sqrt{n})$.*

2. Overview and Discussion

We begin with an outline of our results and techniques in Section 2.1. In Section 2.2 we contrast our techniques to other closely related works for this problem. Lastly, Section 2.3 gives a brief overview of other related works and problems.

2.1. Proof Overview

To keep the discussion simple, we illustrate our ideas in the simplest non-trivial setting where the planted subgraph is a bipartite graph i.e., the case of two color classes ($r = 2$) Our main technical contribution is that we introduce a novel SDP relaxation, the 2-COL SDP which "searches" for both large independent sets of the planted bipartite graph S i.e. S_1 and S_2 simultaneously.

Simultaneous SDP Relaxation. We consider the SDP relaxation for the case of $r = 2$ below,

2-COL SDP (2-Coloring SDP)

$$\max \quad \sum_{i \in V} \langle \mathbf{v}_i, \mathbf{u}_1 + \mathbf{u}_2 \rangle$$

subject to

$$\begin{aligned}
\langle \mathbf{v}_i, \mathbf{v}_j \rangle &= 0 && \forall \{i,j\} \in E \\
\langle \mathbf{v}_i, \mathbf{u}_1 + \mathbf{u}_2 \rangle &= \|\mathbf{v}_i\|^2 && \forall i \in V && (1)\\
\langle \mathbf{v}_i, \mathbf{u}_1 + \mathbf{u}_2 \rangle &\leq 1 && \forall i \in V && (2)\\
\langle \mathbf{u}_1, \mathbf{u}_2 \rangle &= 0 && && (3)\\
\|\mathbf{u}_1\|^2 &= 1 && && (4)\\
\|\mathbf{u}_2\|^2 &= 1 && && (5)
\end{aligned}$$

The "intended" integer solution to 2-COL SDP (see Fig. 2) assigns a vector $\mathbf{e}_1$ to the vector $\mathbf{u}_1$ and vectors corresponding to all the vertices in the independent set S_1, assigns the vector $\mathbf{e}_2$ to $\mathbf{u}_2$ and vectors corresponding to the vertices in S_2 and the zero vector $\mathbf{0}$ to the vectors corresponding to vertices in $V \setminus S$ (here $\mathbf{e}_1$ and $\mathbf{e}_2$ are the standard basis vectors). It is easy to verify that this intended solution satisfies all constraints and is a feasible solution for 2-COL SDP. One can then use the intended solution to obtain a lower bound on the objective value of 2-COL SDP as,

$$\begin{aligned}
\mathrm{opt}\,(\text{2-COL SDP}) &= \sum_{i \in V} \langle \mathbf{v}_i, \mathbf{u}_1 + \mathbf{u}_2 \rangle \geq \sum_{i \in S} \langle \mathbf{v}_i, \mathbf{e}_1 + \mathbf{e}_2 \rangle \\
&= \sum_{i \in S_1} \langle \mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2 \rangle + \sum_{i \in S_2} \langle \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2 \rangle = k. \qquad (6)
\end{aligned}$$

Our 2-COL SDP draws inspiration from the SDP formulation for Lovász Theta Function due to Lovasz (1979), and its planted clique variant considered in the work Feige and Krauthgamer (2000). The novelty in our SDP is that it "searches" for the vertices of all color classes simultaneously. This is in contrast to previous related works for this problem which primarily focused on identifying the edges within the color classes of the graph (see Section 2.2 for a discussion). Another notable feature of our SDP is the constraint $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for adjacent vertices, rather than the conventional r-coloring SDP constraint $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = -1/(r-1)$. This constraint extends the maximum independent set SDP naturally to the setting of r-colorable graphs, by treating each color class as a separate independent set. The constraint $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ and constraint that $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = 0$ ensures that our SDP "tries" to cluster the vectors of each color class such that they are pairwise orthogonal. Especially the constraint involving $\mathbf{u}_1, \mathbf{u}_2$ impart a pronounced geometric structure to our SDP (akin to "crude" SDP considered in the works of Kolla et al. (2011); Makarychev et al. (2012); McKenzie et al. (2020)). We also note that our SDP is not a relaxation of a standard $0-1$ program (due to presence of auxiliary vectors $\mathbf{u}_1, \mathbf{u}_2$ and the constraint $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = 0$. In other words, there is no natural quadratic program for which our SDP can be thought of as relaxation to.

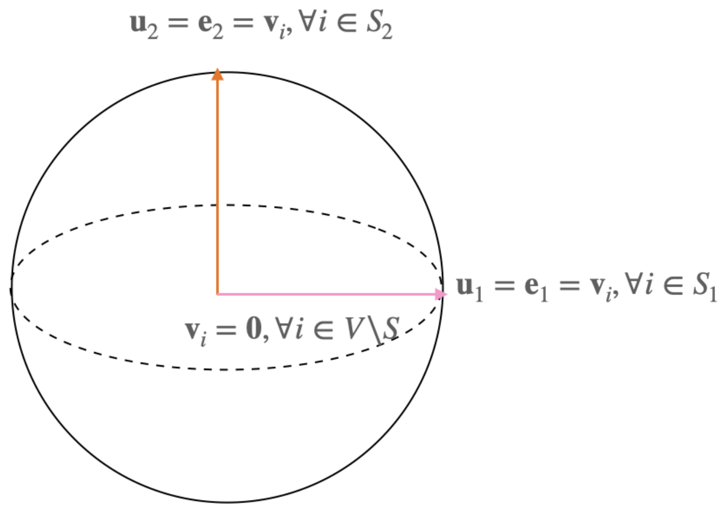

Figure 2: Intended Integral Solution to 2-COL SDP

SDP Analysis. As established above in Eq. (6), the optimal value of the SDP is at least k. Since the objective function is a linear function over the vertices of the graph (instead of being over the edges), a clean decomposition of contributions from S and $V \setminus S$ becomes possible. The analysis then proceeds by bounding the contribution to the objective function from vertices outside the planted set i.e., $V \setminus S$. Specifically, if one can show that the optimal value of the SDP restricted to the graph induced on $V \setminus S$ is $\mathcal{O}(\sqrt{n})$, then one would obtain that the contribution from S in an optimal solution to the SDP (denoted by vectors $\{\mathbf{u}_1^*, \mathbf{u}_2^*, \{\mathbf{v}_i^*\}_{i \in V}\}$) on the whole graph should be,

$$\sum_{i \in S} \langle \mathbf{v}_i^*, \mathbf{u}_1^* + \mathbf{u}_2^* \rangle = \sum_{i \in V} \langle \mathbf{v}_i^*, \mathbf{u}_1^* + \mathbf{u}_2^* \rangle - \sum_{i \in V \setminus S} \langle \mathbf{v}_i^*, \mathbf{u}_1^* + \mathbf{u}_2^* \rangle \geq k - \mathcal{O}\left(\sqrt{n}\right)$$

If k is sufficiently larger than $\sqrt{n}$, we can guarantee that this contribution from planted set S is,

$$\sum_{i \in S} \langle \mathbf{v}_i^*, \mathbf{u}_1^* + \mathbf{u}_2^* \rangle \geq k - \mathcal{O}\left(\sqrt{n}\right) \geq (1-\delta)\, k \text{ for a small constant } \delta > 0.$$

Thus one obtains that majority of the SDP objective mass comes from the set S.[1] Now using the constraint (1), the SDP mass equals $\sum_{i\in V}\|\mathbf{v}_i^*\|^2$. The constraint (2) implies that each vector has length at most 1. Now since $|S|=k$, together with large SDP mass $\sum_{i\in S}\|\mathbf{v}_i^*\|^2\approx k$, it implies that for most vertices $i\in S$ must have $\|\mathbf{v}_i^*\|^2$ close to 1. Consequently, a simple threshold rounding that selects the vertices whose corresponding vectors that have a squared length close to 1 will recover most of the planted set S. Therefore, all that remains to be proven is that the SDP value for random graphs is $\mathcal{O}\left(\sqrt{n}\right)$.

Towards this we introduce another novel idea of bounding the cost of our 2-COL SDP by the standard Lovász Theta function value $\vartheta(G)$.

Lovász Theta SDP

$$\max \quad \left\|\sum_{i\in V}\mathbf{w}_i\right\|^2$$

subject to

$$\langle \mathbf{w}_i,\mathbf{w}_j\rangle = 0 \qquad \forall\{i,j\}\in E$$
$$\sum_{i\in V}\|\mathbf{w}_i\|^2 = 1$$

Note that $\vartheta(G)\overset{\text{def}}{=}\mathsf{opt}(\text{Lovász Theta SDP})$. We can show that after a suitable scaling, the vectors of 2-COL SDP form a feasible solution to the Lovász Theta SDP. We can use this to bound our SDP optimal solution in terms of $\vartheta(G)$ as,

Lemma 7 *For a graph G, the optimal value of 2-COL SDP satisfies,* $\mathsf{opt}\left(\text{2-COL SDP}\right)\le 2\vartheta(G)$.

Therefore, the contribution to the objective value of 2-COL SDP from vertices in $V\setminus S$ can be bounded by $2\vartheta(G[V\setminus S])$. Note that $G[V\setminus S]$ is a random graph $G(n',p)$ where $n'=n-k$, and we can invoke known results, such as Juhász (1982), which show $\vartheta(G(n',p))=\mathcal{O}\left(\sqrt{n'/p}\right)$ with high probability (under suitable conditions on p). This yields the desired bound of $\mathcal{O}\left(\sqrt{n/p}\right)$ on the SDP contribution from $V\setminus S$, completing the proof sketch. The rest of our overview will focus on proving Lemma 7.

Proof [Proof of Lemma 7] Let $\left\{\{\mathbf{v}_i^*\}_{i\in V},\mathbf{u}_1^*,\mathbf{u}_2^*\right\}$ be an optimal solution to 2-COL SDP. Let $\mathsf{optval}\overset{\text{def}}{=}\mathsf{opt}\left(\text{2-COL SDP}\right)$. By constraint (1), it follows that $\mathsf{optval}=\sum_{i\in V}\|\mathbf{v}_i^*\|^2$. Assume $\mathsf{optval}>0$. Define $\mathbf{w}_i=\mathbf{v}_i^*/\sqrt{\mathsf{optval}}$ for all $i\in V$. We now proceed to verify feasibility of $\{\mathbf{w}_i\}_{i\in V}$ for Lovász Theta SDP by noting that,

- For $\{i,j\}\in E$, $\langle\mathbf{w}_i,\mathbf{w}_j\rangle=\left\langle\mathbf{v}_i^*,\mathbf{v}_j^*\right\rangle/\mathsf{optval}=0/\mathsf{optval}=0$.
- $\sum_{i\in V}\|\mathbf{w}_i\|^2=\sum_{i\in V}\|\mathbf{v}_i^*\|^2/\mathsf{optval}=\mathsf{optval}/\mathsf{optval}=1$.

1. Note that if the objective function were defined over the edges of graph instead of the vertices, then in addition to bounding the contribution from $V\setminus S$, we would also have to bound the contribution from the edges between S and $V\setminus S$; in our SDP formulation we do not have the latter term.

Thus, $\{\mathbf{w}_i\}_{i\in V}$ are feasible for Lovász Theta SDP, implying $\left\|\sum_{i\in V}\mathbf{w}_i\right\|^2 \le \vartheta(G)$. Next, consider the objective function value optval and note that,

$$
\begin{aligned}
\mathsf{optval} &= \sum_{i\in V}\langle \mathbf{v}_i^*, \mathbf{u}_1^* + \mathbf{u}_2^*\rangle = \left\langle \sum_{i\in V}\mathbf{v}_i^*, \mathbf{u}_1^* + \mathbf{u}_2^*\right\rangle \\
&\le \left\|\sum_{i\in V}\mathbf{v}_i^*\right\| \|\mathbf{u}_1^* + \mathbf{u}_2^*\| && \text{(By Cauchy-Schwarz)} \\
&= \left\|\sum_{i\in V}\mathbf{v}_i^*\right\| \sqrt{\|\mathbf{u}_1^*\|^2 + \|\mathbf{u}_2^*\|^2 + 2\langle \mathbf{u}_1^*, \mathbf{u}_2^*\rangle} && \text{(Using SDP constraints (4) and (5))} \\
&= \left\|\sum_{i\in V}\mathbf{v}_i^*\right\| \sqrt{1+1+0} = \sqrt{2}\left\|\sum_{i\in V}\mathbf{v}_i^*\right\| && \text{(Using SDP constraints (3))} \\
&= \sqrt{2}\left\|\sum_{i\in V}\mathbf{w}_i\sqrt{\mathsf{optval}}\right\| = \sqrt{2\mathsf{optval}}\left\|\sum_{i\in V}\mathbf{w}_i\right\| && \text{(By construction of } \mathbf{w}_i\text{)} \\
&= \sqrt{2\mathsf{optval}}\sqrt{\left\|\sum_{i\in V}\mathbf{w}_i\right\|^2} \\
&\le \sqrt{2\mathsf{optval}}\sqrt{\vartheta(G)} && \left(\text{Since } \left\|\sum \mathbf{w}_i\right\|^2 \le \vartheta(G)\right)
\end{aligned}
$$

We thus have that optval $\le \sqrt{2\mathsf{optval}\,\vartheta(G)}$. Since we assumed optval > 0, we can divide by $\sqrt{\mathsf{optval}}$ on both sides to get $\sqrt{\mathsf{optval}} \le \sqrt{2\vartheta(G)}$. Squaring both sides yields optval $\le 2\vartheta(G)$. If optval $= 0$, the inequality holds trivially and we get the desired result. ∎

Recovering Planted r-colorable subgraphs. The framework although discussed for 2-COL SDP, extends naturally to the case of recovering planted r-colorable subgraphs for $r \ge 3$. The SDP (see r-COL SDP in Appendix A) can be generalized to use r mutually orthogonal unit vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$. The core analysis strategy remains analogous: establish a lower bound based on the planted solution, relate the SDP value to $\vartheta(G)$ (as shown generally in Lemma 15), bound the contribution from the non-planted part using random graph properties of $\vartheta(G)$, and apply threshold rounding. We refer to Appendix A for the complete details.

Robustness and Exact Recovery. Our algorithm is robust to the presence of a *strong* adversary that can inspect the random graph and insert arbitrary edges in the subgraph induced on $S \times (V \setminus S)$. This follows from the fact that our SDP formulation and analysis do not rely on the structure of these edges. Moreover, it is easy to see (refer Corollary 10 for details) to that the SDP can also tolerate a monotone adversary that can add edges in the subgraph induced on $V \setminus S$.

We also show in Corollary 19 that in the model considered by Kumar et al. (2022) (where $r = 2$ and planted subgraph S is d-regular), one can use our algorithm to achieve exact recovery (for suitable choices of n, k, p). This is achieved by combining our SDP based algorithm with a refining step from Ghoshal et al. (2019); Kumar et al. (2022) that counts the size of maximum matching in the neighborhood subgraph (originally proposed by Alon and Kahalé (1997)).

2.2. Comparison to closely related works

Blackbox MIS Algorithms A natural baseline approach could be to use existing algorithms for the maximum independent set (MIS) in graphs, since the subgraph induced on S consists of r independent sets. There is a rich theory of algorithms for the maximum independent set in random and semirandom instances Alon et al. (1998); Feige and Krauthgamer (2000); Feige and Kilian (2001); McKenzie et al. (2020); Buhai et al. (2023); Błasiok et al. (2024). For instance, when $r = 2$, we could let $S' = S_1$ (or S_2), then the remaining vertices are $V \setminus S' = S_2 \cup (V \setminus S)$ and now one could try to recover S' using the existing results for MIS. However, this approach runs into problems since the edges incident on the large independent set S_1 (or S_2) are not purely random, but have an arbitrary component. Moreover, even if we do find a large independent set, which could

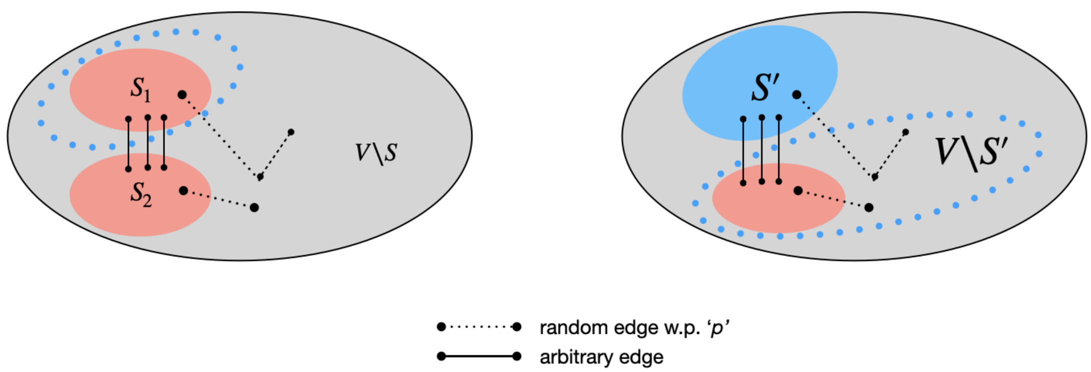

Figure 3: Running Blackbox MIS Algorithms

contain vertices from S_1, S_2 and $V \setminus S$, it is not immediately clear how to recover the rest of S from this since our input subgraph induced on S can be arbitrary 2-colorable graph.

This is the crux of the matter: the algorithm must handle arbitrary edges between color classes (here S_1 has random edges to $V \setminus S$ but arbitrary edges to S_2). It is worth noting that even state-of-the-art algorithms for planted Maximum Independent Set (MIS), such as the works of Buhai et al. (2023); Błasiok et al. (2024), do not provide guarantees in the presence of arbitrary edges at the boundary of the planted set S' (which is S_1 in Fig. 3).

Greedy Algorithms. In the planted clique/MIS problem, there is a naive greedy algorithm that recovers the planted set (w.h.p.) by simply picking out vertices with large/small degrees when k becomes as large as $\Omega(\sqrt{n \log n})$. However, if we set the average degree d of the planted set S as $d = pk$, such naive tests are no longer feasible for the planted arbitrary r-colorable subgraphs, even for a large k such as $k \geq cn$ (for some constant $c \leq 1$). This is because the expected degrees of planted and non-planted vertices is indistinguishable as for a vertex $u \in S, v \notin S$,

$$\mathbb{E}[\deg(u)] = d + p(n-k) = pk + p(n-k) = pn = \mathbb{E}[\deg(v)].$$

In a recent breakthrough work Błasiok et al. (2024), the authors propose a powerful greedy algorithm for planted clique problem that recovers for $k = \Omega(\sqrt{n} \log^2 n)$ in a model that has a strong adversarial presence (subsumes the random planted model for clique problem). Their algorithm is based on repeated tensoring of the adjacency matrix and demonstrating certain Restricted Isometry Property (RIP) properties for the same. However, as such their techniques will also fail when $d \approx pk$, the regimes of interest in our setting.

In fact there are algorithms for the more general densest k-subgraph problem Bhaskara et al. (2010); Hajek et al. (2016b,c,d) that can recover an average degree d planted graph but again these fail to recover in $d \approx pk$ regimes. In contrast our SDP based algorithm continues to work in these regimes. This underscores the novelty of our SDP approach which is able to sidestep this issue by switching to the objective value based on the vertices of the graph instead of edges of the graph.

Spectral Approaches. Consider the detection version of the problem where the task is to distinguish the case when the given graph contains a planted r-colorable subgraph of average degree d, from the case when the input graph is a random graph. Formally, we have the following two scenarios:

- $\mathcal{H}_1 : G \sim G(n, k, p, d)$, a planted r-colorable graph as in Definition 1 against
- $\mathcal{H}_0 \sim G(n, p)$, an Erdős-Rényi random graph.

In case the planted r-coloring is also random, the works Krivelevich (2002); Coja-Oghlan (2005) give a distinguishing algorithm for $k = \Omega(r\sqrt{n})$, while the work Kothari et al. (2023) rules out any distinguishing test for $k = o(r\sqrt{n})$ (assuming the Planted Clique Hypothesis). Even if the planted r-colorable subgraph is arbitrary as in Definition 1 and has a sufficiently large average degree $d = \Omega(r\sqrt{pn})$, a simple test that examines the smallest eigenvalue can distinguish it from a random graph (see Appendix B for details). We note that our SDP Recovery Algorithm and in our corresponding guarantees in Theorem 3, we don't make any such density assumptions and hence we can also use the algorithm to distinguish the two hypothesis for $k = \Omega_p(r\sqrt{n})$

For comparison, recall the planted clique setting where an indicator vector $\mathbf{1}_S$ achieves a Rayleigh quotient of value $k - 1$ which dominates the $2\sqrt{pn}$ contribution from a random graph on $V \setminus S$ and allows detecting a planted clique for $k = \Omega(\sqrt{n})$. Further, the work Alon et al. (1998) showed that this spectral test also leads to a spectral algorithm based on examining an appropriate eigenvector of the graph, and they show how to recover the planted clique. Therefore, it is natural to ask can we turn the simple spectral distinguishing test discussed above into a spectral algorithm for recovering the planted r-colorable graph.

To this end, one might conjecture that the bottom few eigenvectors should be "correlated" with the indicator vector of the planted subset S. However, proving this is unlikely to be easy since little is known about the structure of the bottom eigenvectors of a general r-colorable graphs. In fact, even our understanding of the top eigenvector in a graph with average degree d is quite limited.

The work Kumar et al. (2022) presents a spectral algorithm that recovers an arbitrary planted r-coloring when $k = \Omega_p(\sqrt{n})$, but only for $r = 2$ and with an additional assumption that S is d-regular bipartite graph. Their algorithm runs in time $n^{\mathcal{O}(1/p)}$,which is not a polynomial time algorithm in the sparse regime when $p = o(1)$. Their algorithm is based on their observation that the signed indicator vector of the S (i.e. $\mathbf{1}_{S_1} - \mathbf{1}_{S_2}$) has a large projection on the subspace spanned by eigenvectors having eigenvalue close to $-d$. They then use "subspace enumeration" ideas of Kolla and Tulsiani (2007); Arora et al. (2010); Kolla (2011) over this subspace to approximately recover the signed indicator vector of S. It is however unclear if these ideas can be extended to r-colorable graphs or even for an average degree d bipartite graphs ($r = 2$ case).

However, if the planted r-colorable graph is itself random, we expect that the subspace enumeration approach along lines of Kumar et al. (2017, 2022) could efficiently recover the planted r-colorable graph for $k = \Omega_p(r\sqrt{n})$.

SDP Relaxations. An alternative approach involves uses Semidefinite Programming. Consider the following SDP relaxation for case of $r = 2$ proposed in Kumar et al. (2022):

Spectral SDP

$$\min \sum_{\{i,j\}\in E} 2\langle \mathbf{y}_i, \mathbf{y}_j \rangle$$

subject to

$$\sum_{i\in V} \|\mathbf{y}_i\|^2 = k$$

$$\|\mathbf{y}_i\|^2 \leq 1 \qquad \forall i \in V$$

$$\langle \mathbf{y}_i, \mathbf{y}_j \rangle \leq 0 \qquad \forall \{i,j\} \in E$$

Here, the intended integral solution to Spectral SDP corresponds to planted bipartition as,

$$\mathbf{y}_i = \begin{cases} +\mathbf{e} & \forall i \in S_1 \\ -\mathbf{e} & \forall i \in S_2 \\ \mathbf{0} & \forall i \in V \setminus S \end{cases} \tag{7}$$

where $\mathbf{e}$ is a unit vector. The work Kumar et al. (2022) analyzes the Spectral SDP and gives a polynomial-time algorithm to recover S when $k = \Omega_p\left(\sqrt{n \log n}\right)$. At a high level, their algorithm involves constructing a high rank dual solution to show integrality of Spectral SDP. They construct the high rank dual solution by handcrafting the value of dual variables so that the eigenvectors of adjacency matrix (after a suitable shift) also serve as the eigenvectors of the dual SDP matrix. They call this as the *calibrating eigenvectors framework*, and we refer to Section 1.4 of Kumar et al. (2022) for more details.

We recall again that analysis by Kumar et al. (2022) is only for d-regular graphs. However, for average degree d graphs, Spectral SDP is unlikely to be able to recover the entire planted set S. This is since the objective function is solely supported on the edges of the graph, and therefore the sparse regions of planted subgraph may contribute little to the objective. Consider the following instance of a planted bipartite graph:

Construction 8 *Let $S = S_1 \cup S_2$ be the vertex set of the planted bipartite graph. Suppose there exist subsets $S_1' \subset S_1$ and $S_2' \subset S_2$ such that $G[S_1' \cup S_2']$ is relatively dense, while the induced subgraph $G[(S_1 \setminus S_1') \cup (S_2 \setminus S_2')]$ is very sparse or even an independent set. Furthermore, assume few or no edges connect $(S_1 \setminus S_1')$ to S_2' or $(S_2 \setminus S_2')$ to S_1'.*

If there are no edges in the subgraph induced on $(S_1 \setminus S_1') \cup (S_2 \setminus S_2')$, Spectral SDP doesn't gain anything in the objective value by putting SDP mass here and instead would be better off putting this SDP mass on $V \setminus S$. Even if the subgraph induced on $(S_1 \setminus S_1') \cup (S_2 \setminus S_2')$ is not empty, the SDP might achieve a better (lower) objective value by setting $\mathbf{y}_i = \mathbf{0}$ for $i \in (S \setminus S')$ and distributing the total mass $\sum \|\mathbf{y}_i\|^2 = k$ primarily across $S' = S'_1 \cup S'_2$ and potentially some vertices in $V \setminus S$ that have favorable edge connections (or lack thereof) relative to S'.

2.3. Other related works

Planted Random Models and Semirandom Models. Planted Random models and Semirandom models has recently been a popular framework in study of robust algorithm design for computationally hard optimization problems. As noted earlier, there are combinatorial algorithms Kucera (1995); Feige and Ron (2010); Dekel et al. (2014), and spectral algorithms Alon and Kahalé (1997); Alon et al. (1998); McSherry (2001) that recover planted graphs in random models. There are host of works based on SDP relaxations for the planted clique/independent set problem Feige and Krauthgamer (2000); Feige and Kilian (2001); McKenzie et al. (2020); Buhai et al. (2023), the planted dense subgraph problem Bhaskara et al. (2010); Hajek et al. (2016b,c,d), Stochastic Block Models Feige and Kilian (2001); Abbe and Sandon (2015); Chen and Xu (2016); Abbe et al. (2016), and more generally, graph partitioning problems Makarychev et al. (2012, 2014); Louis and Venkat (2018, 2019). The SDP based algorithms have an additional advantage that they are robust to monotone adversaries and can also recover the planted set S in semirandom models. We refer the reader to Chapter 10 in Roughgarden (2021) for a discussion on robustness of SDPs to monotone adversaries.

Graph Coloring. In the classical graph coloring problem, given a graph $G = (V, E)$, the goal is to assign colors to the vertices of the graph with as few colors as possible, such that for every edge $\{u, v\} \in E$, the endpoints u and v receive different colors. It is one of the Karp's list Karp (1972) of 21 original NP-hard problems. Further, Feige and Kilian (1998); Zuckerman (2007) showed that it is NP-hard to approximate better than $n^{1-\varepsilon}$ for any constant $\varepsilon > 0$.

Consequently, a lot of the focus has been in studying the r-coloring problem (largely $r = 3$). In the r-coloring problem, we are promised that the graph is r-colorable and the aim is still to color the graph with minimum number of colors. The problem was also shown to be NP-hard by Karp (1972). There is also a rich literature of algorithms (mostly SDP based) for 3-colorable graphs due to Wigderson (1983); Blum (1994); Karger et al. (1998); Arora et al. (2006); Chlamtac (2007); Kawarabayashi and Thorup (2012, 2017) with the current best known algorithm due to Kawarabayashi et al. (2024) which gives $\mathcal{O}\left(n^{0.197}\right)$-approximation.

Random Planted Colorings. There has also been keen interest in random planted instances of the r-coloring problem. Starting with the works of Dyer and Frieze (1989); Blum and Spencer (1995); Coja-Oghlan (2004) who studied the problem in $G(n, p, r)$ model where the vertex set is partitioned into r color classes and edges between color classes are added independently with probability p. The work by Alon and Kahalé (1997) gave a spectral algorithm for $r = 3$ that works for $p \geq c/n$ (for some constant c). Planted models for this problem were also considered in the *hosted coloring framework* by David and Feige (2016). They extend the spectral algorithm of Alon and Kahalé (1997) and adapt it to the setting when a random coloring is planted on a d-regular spectral expander.

Recently, Kothari et al. (2023) showed that recovering a r-coloring in $G(n, 1/2, r)$ model (r color classes of equal size say k) is as hard as recovering clique in the low degree polynomial model of computation, and thus we need each color class to be of size at least $\Omega\left(\sqrt{n}\right)$ for efficient recovery to be possible. They also note that distinguishing the planted coloring $\mathcal{H}_1 \sim G(n, 1/2, r)$ and $\mathcal{H}_0 \sim G_{n,p}$ is easy for $k \gg \sqrt{n}$ by simply counting the total number of edges. For the hardness result, they consider more quieter null distributions which is a $q+1$ colorable graph and show hardness of distinguishing these for $k = o(\sqrt{n})$ in the low degree polynomial framework. They essentially argue that finding the last color hidden color class is hard, even if an oracle revealed all the others to you as it is still an instance of the planted clique problem.

Partial Coloring Problem. Another line of work by Kumar et al. (2017); Ghoshal et al. (2019), considers the *partial coloring problem* where a large fraction of the graph is r-colorable. They note that this is a more robust structural property than strict r-colorability and hence more useful for graphs arising in practice. They study both the worst-case instances and some "strong" semirandom instances of the problem. The work Ghoshal et al. (2019) also give a partial recovery algorithm, but for the regimes of $k = (1-\varepsilon)n$ (for small ε). Our models in Definition 1 and Definition 2 can also be thought of as planted and semirandom versions of the partial coloring problem respectively.

Lower Bounds. The work Kothari et al. (2023) also consider the planted partial coloring problem. They study the problem in the special case where the edges between the color classes in Definition 1 are added randomly instead of arbitrarily. Further they assume that each color class has the same size $|S_i| = k/r$, and focus on the case when $p = 1/2$. They show that a simple degree counting algorithm (similar to Kucera (1995)) can fully recover the planted graph w.h.p. for $k = \Omega\left(r\sqrt{n\log n}\right)$ regimes. Their main result is a lower bound of $k = o(r\sqrt{n})$ assuming the Planted Clique Hypothesis. Since their model makes stronger assumptions, the lower bound also applies to our models in Definition 1 and Definition 2. Additionally, they show that distinguishing the partial planted model from the null hypothesis of random graphs $\mathcal{H}_0 : G_{n,1/2}$ is hard for $k = o(\sqrt{n})$ by a reduction to finding a standard planted clique.

The work Brennan et al. (2018) considers a collection of sparse planted problems and show sharp lower bounds for transition between computationally hard and easy regimes for these problems. Also the lower bounds of statistical query algorithms Feldman et al. (2017) and SoS algorithms Barak et al. (2019) apply here. However, this is not the focus of our work and we make no attempts to optimize the constants and parameters p and δ in our Theorem 3.

Other Models for Beyond Worstcase Analysis. Beyond planted and semirandom models, several alternative frameworks have been proposed for designing algorithms beyond worst-case. These include perturbation resilient instances, paramterized algorithms, smoothed analysis, and self-improving algorithms; see Roughgarden (2021) for a comprehensive overview. While it is beyond the scope of this paper to cover each of these approaches, we talk about one particularly relevant approach of *low threshold rank graphs*.

This approach was pioneered by the works of Arora et al. (2010); Barak et al. (2011); Guruswami and Sinop (2011) who studied various optimization problems on such low rank graphs. These are graphs whose random walk matrix have all but small number of eigenvalues bounded far from a fixed threshold τ. The choice of τ is tailored to the problem at hand, close to $+1$ for partitioning problems and close to -1 for coloring and independent set problems. The work by Arora and Ge (2011) studies the r-coloring problem (for $r = 3$) on low rank graphs for choice of $\tau = -(1-\varepsilon)$. They show that for any constant $\varepsilon > 0$ and a graph with rank at most k, an algorithm can return an independent set of size cn (where c is a constant) using a $\mathcal{O}(k)$-level Lasserre (Sum-of-Squares) relaxation.

Acknowledgments

AL was supported in part by the Walmart Center for Tech Excellence at IISc, and SERB award CRG/2023/002896. AL and PR were supported in part by the IUSSTF virtual center on "Polynomials as an Algorithmic Paradigm". RP was supported by Prime Minister's Research Fellowship, India. We thank Santosh Vempala for helpful discussions. We thank the anonymous reviewers for their comments on an earlier version of this paper.

References

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 670–688. IEEE Computer Soc., Los Alamitos, CA, 2015. doi: 10.1109/FOCS.2015.47. URL https://doi.org/10.1109/FOCS.2015.47.

Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory*, 62(1):471–487, 2016. ISSN 0018-9448. doi: 10.1109/TIT.2015.2490670. URL https://doi.org/10.1109/TIT.2015.2490670.

Noga Alon and Nabil Kahalé. A spectral technique for coloring random 3-colorable graphs. *SIAM J. Comput.*, 26(6):1733–1748, 1997.

Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*, volume 13, pages 457–466, 1998. doi: 10.1002/(SICI)1098-2418(199810/12)13:3/4⟨457::AID-RSA14⟩3.3.CO;2-K. URL https://doi.org/10.1002/(SICI)1098-2418(199810/12)13:3/4<457::AID-RSA14>3.3.CO;2-K.

Sanjeev Arora and Rong Ge. New tools for graph coloring. In *Approximation, randomization, and combinatorial optimization*, volume 6845 of *Lecture Notes in Comput. Sci.*, pages 1–12. Springer, Heidelberg, 2011. doi: 10.1007/978-3-642-22935-0_1. URL https://doi.org/10.1007/978-3-642-22935-0_1.

Sanjeev Arora, Eden Chlamtac, and Moses Charikar. New approximation guarantee for chromatic number. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 215–224. ACM, New York, 2006. doi: 10.1145/1132516.1132548. URL https://doi.org/10.1145/1132516.1132548.

Sanjeev Arora, Boaz Barak, and David Steurer. Subexponential algorithms for unique games and related problems. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science—FOCS 2010*, pages 563–572. IEEE Computer Soc., Los Alamitos, CA, 2010.

Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011*, pages 472–481. IEEE Computer Soc., Los Alamitos, CA, 2011. doi: 10.1109/FOCS.2011.95. URL https://doi.org/10.1109/FOCS.2011.95.

Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K. Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019. doi: 10.1137/17M1138236. URL https://doi.org/10.1137/17M1138236.

Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities—an $O(n^{1/4})$ approximation for densest k-subgraph. In *STOC'10—Proceedings of the 2010 ACM International Symposium on Theory of Computing*, pages 201–210. ACM, New York, 2010.

Avrim Blum. New approximation algorithms for graph coloring. *J. Assoc. Comput. Mach.*, 41(3): 470–516, 1994. ISSN 0004-5411. doi: 10.1145/176584.176586. URL https://doi.org/10.1145/176584.176586.

Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *J. Algorithms*, 19(2):204–234, 1995.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 48–166. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/brennan18a.html.

Rares-Darius Buhai, Pravesh K. Kothari, and David Steurer. Algorithms approaching the threshold for semi-random planted clique. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1918–1926, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585184. URL https://doi.org/10.1145/3564246.3585184.

Jarosław Błasiok, Rares-Darius Buhai, Pravesh K. Kothari, and David Steurer. Semirandom planted clique and the restricted isometry property, 2024. URL https://arxiv.org/abs/2404.14159.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.*, 17: Paper No. 27, 57, 2016. ISSN 1532-4435.

Eden Chlamtac. Approximation algorithms using hierarchies of semidefinite programming relaxations. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 691–701, 2007. doi: 10.1109/FOCS.2007.72.

Amin Coja-Oghlan. Coloring semirandom graphs optimally. In Josep Díaz, Juhani Karhumäki, Arto Lepistö, and Donald Sannella, editors, *Automata, Languages and Programming*, pages 383–395, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-27836-8.

Amin Coja-Oghlan. The Lovász number of random graphs. *Combin. Probab. Comput.*, 14(4):439–465, 2005. ISSN 0963-5483. doi: 10.1017/S0963548305006826. URL https://doi.org/10.1017/S0963548305006826.

Roee David and Uriel Feige. On the effect of randomness on planted 3-coloring models. In *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 77–90. ACM, New York, 2016. doi: 10.1145/2897518.2897561. URL https://doi.org/10.1145/2897518.2897561.

Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. *Combin. Probab. Comput.*, 23(1):29–49, 2014. ISSN 0963-5483. doi: 10.1017/S096354831300045X. URL https://doi.org/10.1017/S096354831300045X.

Yash Deshpande and Andrea Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Found. Comput. Math.*, 15(4):1069–1128, 2015. ISSN 1615-3375. doi: 10.1007/s10208-014-9215-y. URL https://doi.org/10.1007/s10208-014-9215-y.

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

Uriel Feige and Vadim Grinberg. How to hide a clique? In *47th International Colloquium on Automata, Languages, and Programming*, volume 168 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 44, 13. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2020.

Uriel Feige and Joe Kilian. Zero knowledge and the chromatic number. volume 57, pages 187–199. 1998. doi: 10.1006/jcss.1998.1587. URL https://doi.org/10.1006/jcss.1998.1587. Complexity 96—The Eleventh Annual IEEE Conference on Computational Complexity (Philadelphia, PA).

Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computing and System Sciences*, 63:639–671, 2001.

Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures Algorithms*, 16(2):195–208, 2000. ISSN 1042-9832. doi: 10.1002/(SICI)1098-2418(200003)16:2⟨195::AID-RSA5⟩3.3.CO;2-1. URL https://doi.org/10.1002/(SICI)1098-2418(200003)16:2<195::AID-RSA5>3.3.CO;2-1.

Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, Discrete Math. Theor. Comput. Sci. Proc., AM, pages 189–203. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2010.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2), apr 2017. ISSN 0004-5411. doi: 10.1145/3046674. URL https://doi.org/10.1145/3046674.

Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981. ISSN 0209-9683. doi: 10.1007/BF02579329. URL https://doi.org/10.1007/BF02579329.

Suprovat Ghoshal, Anand Louis, and Rahul Raychaudhury. Approximation algorithms for partially colorable graphs. In *Approximation, randomization, and combinatorial optimization. Algorithms* *and techniques*, volume 145 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 28, 20. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019.

Venkatesan Guruswami and Ali Kemal Sinop. Lasserre hierarchy, higher eigenvalues, and approximation schemes for graphs partitioning and quadratic integer programming with PSD objectives (extended abstract). In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011*, pages 482–491. IEEE Computer Soc., Los Alamitos, CA, 2011. doi: 10.1109/FOCS.2011.36. URL https://doi.org/10.1109/FOCS.2011.36.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: extensions. *IEEE Trans. Inform. Theory*, 62(10):5918–5937, 2016a. ISSN 0018-9448. doi: 10.1109/TIT.2016.2594812. URL https://doi.org/10.1109/TIT.2016.2594812.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inform. Theory*, 62(5):2788–2797, 2016b. ISSN 0018-9448. doi: 10.1109/TIT.2016.2546280. URL https://doi.org/10.1109/TIT.2016.2546280.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: extensions. *IEEE Trans. Inform. Theory*, 62(10):5918–5937, 2016c. ISSN 0018-9448. doi: 10.1109/TIT.2016.2594812. URL https://doi.org/10.1109/TIT.2016.2594812.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Semidefinite programs for exact recovery of a hidden community, 2016d.

A. J. Hoffman and Leonard Howes. On eigenvalues and colorings of graphs. II. *Ann. New York Acad. Sci.*, 175:238–242, 1970. ISSN 0077-8923.

Alan J. Hoffman. On eigenvalues and colorings of graphs. In *Graph Theory and its Applications (Proc. Advanced Sem., Math. Research Center, Univ. of Wisconsin, Madison, Wis., 1969)*, pages 79–91. Academic Press, New York-London, 1970.

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013. ISBN 978-0-521-54823-6.

Mark Jerrum. Large cliques elude the Metropolis process. *Random Structures Algorithms*, 3(4): 347–359, 1992. ISSN 1042-9832. doi: 10.1002/rsa.3240030402. URL https://doi.org/10.1002/rsa.3240030402.

Ferenc Juhász. The asymptotic behaviour of Lovász' θ function for random graphs. *Combinatorica*, 2(2):153–155, 1982. ISSN 0209-9683. doi: 10.1007/BF02579314. URL https://doi.org/10.1007/BF02579314.

David Karger, Rajeev Motwani, and Madhu Sudan. Approximate graph coloring by semidefinite programming. *J. ACM*, 45(2):246–265, 1998. ISSN 0004-5411. doi: 10.1145/274787.274791. URL https://doi.org/10.1145/274787.274791.

Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972)*, The IBM Research Symposia Series, pages 85–103. Plenum, New York-London, 1972.

Ken-ichi Kawarabayashi and Mikkel Thorup. Combinatorial coloring of 3-colorable graphs. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science—FOCS 2012*, pages 68–75. IEEE Computer Soc., Los Alamitos, CA, 2012.

Ken-Ichi Kawarabayashi and Mikkel Thorup. Coloring 3-colorable graphs with less than $n^{1/5}$ colors. *J. ACM*, 64(1):Art. 4, 23, 2017. ISSN 0004-5411. doi: 10.1145/3001582. URL https://doi.org/10.1145/3001582.

Ken-ichi Kawarabayashi, Mikkel Thorup, and Hirotaka Yoneda. Better coloring of 3-colorable graphs. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, page 331–339, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703836. doi: 10.1145/3618260.3649768. URL https://doi.org/10.1145/3618260.3649768.

Alexandra Kolla. Spectral algorithms for unique games. *Comput. Complexity*, 20(2):177–206, 2011. ISSN 1016-3328. doi: 10.1007/s00037-011-0011-7. URL https://doi.org/10.1007/s00037-011-0011-7.

Alexandra Kolla and Madhur Tulsiani. Playing random and expanding unique games, 2007.

Alexandra Kolla, Konstantin Makarychev, and Yury Makarychev. How to play unique games against a semi-random adversary: study of semi-random models of unique games. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011*, pages 443–452. IEEE Computer Soc., Los Alamitos, CA, 2011. doi: 10.1109/FOCS.2011.78. URL https://doi.org/10.1109/FOCS.2011.78.

Pravesh Kothari, Santosh S Vempala, Alexander S Wein, and Jeff Xu. Is planted coloring easier than planted clique? In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5343–5372. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/kothari23a.html.

Michael Krivelevich. Deciding k-colorability in expected polynomial time. *Inform. Process. Lett.*, 81(1):1–6, 2002. ISSN 0020-0190. doi: 10.1016/S0020-0190(01)00187-9. URL https://doi.org/10.1016/S0020-0190(01)00187-9.

Ludek Kucera. Expected complexity of graph partitioning problems. *Discret. Appl. Math.*, 57 (2-3):193–212, 1995. doi: 10.1016/0166-218X(94)00103-K. URL https://doi.org/10.1016/0166-218X(94)00103-K.

Akash Kumar, Anand Louis, and Madhur Tulsiani. Finding pseudorandom colorings of pseudorandom graphs. In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 93 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 37, 12. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2017.

Akash Kumar, Anand Louis, and Rameesh Paul. Exact recovery algorithm for planted bipartite graph in semi-random graphs. In *49th EATCS International Conference on Automata, Languages, and Programming*, volume 229 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 84, 20. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2022. doi: 10.4230/lipics.icalp.2022.84. URL https://doi.org/10.4230/lipics.icalp.2022.84.

Anand Louis and Rakesh Venkat. Semi-random graphs with planted sparse vertex cuts: algorithms for exact and approximate recovery. In *45th International Colloquium on Automata, Languages, and Programming*, volume 107 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 101, 15. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.

Anand Louis and Rakesh Venkat. Planted models for k-way edge and vertex expansion. In *39th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 150 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 23, 15. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019.

L. Lovasz. On the shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25(1): 1–7, 1979. doi: 10.1109/TIT.1979.1055985.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In *STOC'12—Proceedings of the 2012 ACM Symposium on Theory of Computing*, pages 367–384. ACM, New York, 2012. doi: 10.1145/2213977.2214013. URL https://doi.org/10.1145/2213977.2214013.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Constant factor approximation for balanced cut in the PIE model. In *STOC'14—Proceedings of the 2014 ACM Symposium on Theory of Computing*, pages 41–49. ACM, New York, 2014.

Cheng Mao, Alexander S. Wein, and Shenduo Zhang. Detection-recovery gap for planted dense cycles. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2440–2481. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/mao23a.html.

Theo McKenzie, Hermish Mehta, and Luca Trevisan. A new algorithm for the robust semi-random independent set problem. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 738–746. SIAM, 2020. doi: 10.1137/1.9781611975994.45. URL https://doi.org/10.1137/1.9781611975994.45.

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

Asaf Rotenberg, Wasim Huleihel, and Ofer Shayevitz. Planted Bipartite Graph Detection. *IEEE Trans. Inform. Theory*, 70(6):4319–4334, 2024. ISSN 0018-9448. doi: 10.1109/tit.2024.3382228. URL https://doi.org/10.1109/tit.2024.3382228.

Tim Roughgarden. *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press, 2021. doi: 10.1017/9781108637435.

Van H. Vu. Spectral norm of random matrices. *Combinatorica*, 27(6):721–736, 2007. ISSN 0209-9683. doi: 10.1007/s00493-007-2190-z. URL https://doi.org/10.1007/s00493-007-2190-z.

Avi Wigderson. Improving the performance guarantee for approximate graph coloring. *J. Assoc. Comput. Mach.*, 30(4):729–735, 1983. ISSN 0004-5411. doi: 10.1145/2157.2158. URL https://doi.org/10.1145/2157.2158.

David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory Comput.*, 3:103–128, 2007. doi: 10.4086/toc.2007.v003a006. URL https://doi.org/10.4086/toc.2007.v003a006.

Appendix A. Recovering Planted r-Colorable Subgraphs

We consider the following SDP (vector program formulation) for our problem which attempts to compute r-colorable subgraphs by finding r separate independent sets simultaneously.

r-COL SDP (r-Coloring SDP)

$$\max \quad \sum_{i \in V} \sum_{l \in [r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle$$

$$\text{subject to} \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \qquad \forall \{i, j\} \in E \tag{8}$$

$$\sum_{l \in [r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle = \|\mathbf{v}_i\|^2 \qquad \forall i \in V \tag{9}$$

$$\sum_{l \in [r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle \leq 1 \qquad \forall i \in V \tag{10}$$

$$\langle \mathbf{u}_l, \mathbf{u}_{l'} \rangle = 0 \qquad \forall l \in [r], \forall l' \in [r] \tag{11}$$

$$\|\mathbf{u}_l\|^2 = 1 \qquad \forall l \in [r] \tag{12}$$

We now present our recovery algorithm based on r-COL SDP

SDP Recovery Algorithm: For an instance of $G = (V, E)$ generated according to Definition 1 for regimes of n, k, r, p, ε satisfying $k \geq \frac{3r}{\varepsilon^2}\sqrt{\frac{n}{p}}$,

1. Solve r-COL SDP. Let $\left\{ \{\mathbf{v}_i^*\}_{i \in V}, \{\mathbf{u}_l^*\}_{l \in [r]} \right\}$ be the optimal solution.

2. Return $T = \left\{ i \in S : \|\mathbf{v}_i^*\|^2 \geq 1 - \varepsilon \right\}$.

Theorem 9 (Formal Version of Theorem 3) *Given a graph $G = (V, E)$ generated according to Definition 1, and for regimes of n, k, r, p, ε satisfying $k \geq \frac{3r}{\varepsilon^2}\sqrt{\frac{n}{p}}$ and $\frac{(\ln n)^7}{n} \leq p \leq \frac{1}{2}$, with high probability (over the randomness of the input), the SDP Recovery Algorithm outputs a set T where,*

- *$|T \cap S| \geq (1-\varepsilon)k$, and*
- *$|T| \leq \left(1 + \frac{\varepsilon^2}{1-\varepsilon}\right) k$.*

Corollary 10 *Given a graph $G = (V, E)$ generated according to Definition 2, and regimes of n, k, r, p, ε satisfying $k \geq \frac{3r}{\varepsilon^2}\sqrt{\frac{n}{p}}$ and $\frac{(\ln n)^7}{n} \leq p \leq \frac{1}{2}$, with high probability (over the randomness of the input instance), the SDP Recovery Algorithm outputs a set T such that,*

- *$|T \cap S| \geq (1-\varepsilon)k$, and*
- *$|T| \leq \left(1 + \frac{\varepsilon^2}{1-\varepsilon}\right) k$.*

Remark 11 *We note that for $p \leq \ln n/n$, the graph induced on $V \setminus S$ is not even connected and has unintentional large independent sets in $V \setminus S$. That said, we have not made any attempts to optimize the dependence on p and ε.*

Lemma 12 *Given an instance of graph $G = (V, E)$ generated according to Definition 1, we have that,*

$$\mathsf{opt}(r\text{-}\mathsf{COL\,SDP}) \geq k.$$

Proof Let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_r$ be unit vectors, which are pairwise orthogonal i.e. $\|\mathbf{e}_l\|^2 = 1, \forall l \in [r]$ and $\langle \mathbf{e}_l, \mathbf{e}_{l'} \rangle = 0, \forall l \in [r], l' \in [r], l \neq l'$. We consider the following intended integral solution for r-COL SDP,

$$\mathbf{v}_i = \begin{cases} \mathbf{e}_1 & \forall i \in S_1 \\ \mathbf{e}_2 & \forall i \in S_2 \\ \vdots & \\ \mathbf{e}_r & \forall i \in S_r \\ \mathbf{0} & \forall i \in V \setminus S \end{cases} \quad \text{and} \quad \mathbf{u}_l = \mathbf{e}_l, \forall l \in [r]. \tag{13}$$

Clearly this satisfies the constraints (8),(11) and (12) by construction. For constraint (9) we can check that it holds if $i \in V \setminus S$ since both sides are 0 and otherwise for $i \in S_t$,

$$1 = \|\mathbf{e}_t\|^2 = \|\mathbf{v}_i\|^2 = \sum_{l \in [r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle = \langle \mathbf{e}_t, \mathbf{e}_t \rangle = 1.$$

and the constraint (10) is true for $i \in V \setminus S$ since the LHS is 0 and also holds for $i \in S_t$ since,

$$\sum_{l \in [r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle = \langle \mathbf{e}_t, \mathbf{e}_t \rangle + \sum_{l \in [r], l \neq t} \langle \mathbf{e}_l, \mathbf{e}_t \rangle = 1.$$

Finally, we compute the objective value of this intended solution and we have that,

$$\begin{aligned}\mathsf{opt}(r\text{-COL SDP}) \geq & \sum_{i\in V}\sum_{l\in[r]}\langle \mathbf{v}_i, \mathbf{u}_l\rangle = \sum_{i\in S}\sum_{l\in[r]}\langle \mathbf{v}_i, \mathbf{u}_l\rangle = \sum_{l\in[r]}\sum_{i\in S}\langle \mathbf{v}_i, \mathbf{u}_l\rangle \\ &= \sum_{l\in[r]}\sum_{i\in S_l}\langle \mathbf{e}_l, \mathbf{e}_l\rangle = \sum_{l\in[r]}|S_l| = k.\end{aligned}$$

■

We consider the SDP relaxation of the classical Lovász Theta function. There are multiple formulations for this given by Lovasz (1979). Here, in Lovász Theta SDP, we consider the face given by Theorem 4 in Lovasz (1979). Let $\vartheta(G)$ denote the optimum value of Lovász Theta SDP.

Lovász Theta SDP

$$\begin{aligned} \max \quad & \left\|\sum_{i\in V}\mathbf{v}_i\right\|^2 & \\ \text{subject to} \quad & \langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0 & \forall\{i,j\}\in E \\ & \sum_{i\in V}\|\mathbf{v}_i\|^2 = 1 & \end{aligned}$$

Definition 13 (Orthogonal Representation of a Graph) *An orthogonal representation of a graph G of n vertices is a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ such that $\langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0, \forall\{i,j\}\in E$.*

Lemma 14 *Given an orthogonal representation of a graph G as $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, we have that,*

$$\left\|\sum_{i\in V}\mathbf{v}_i\right\|^2 \leq \vartheta(G)\sum_{i\in V}\|\mathbf{v}_i\|^2.$$

Proof Given an orthogonal representation of a graph, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, we let $c = \sum_{i\in V}\|\mathbf{v}_i\|^2$. Consider the scaled set of vector $\{\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_n\}$ where $\mathbf{v}'_i = \mathbf{v}_i/\sqrt{c}$. Hence the set of vectors $\{\mathbf{v}'_i\}_{i\in V}$ are a feasible solution to Lovász Theta SDP. Therefore,

$$\vartheta(G) \geq \left\|\sum_{i\in V}\mathbf{v}'_i\right\|^2 = \left\|\frac{\sum_{i\in V}\mathbf{v}_i}{\sqrt{c}}\right\|^2 = \frac{1}{c}\left\|\sum_{i\in V}\mathbf{v}_i\right\|^2 = \frac{\left\|\sum_{i\in V}\mathbf{v}_i\right\|^2}{\sum_{i\in V}\|\mathbf{v}_i\|^2}.$$

. ■

Lemma 15 *For a given graph G we have that,* $\mathsf{opt}(r\text{-COL SDP}) \leq r\vartheta(G)$.

Proof Consider any feasible solution $\left\{ \{\mathbf{v}_i^*\}_{i\in V}, \{\mathbf{u}_l^*\}_{l\in[r]} \right\}$ of r-COL SDP. First we observe using Cauchy-Schwarz inequality and using constraint (11) of r-COL SDP that,

$$\sum_{i\in V}\sum_{l\in[r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle \leq \left\| \sum_{r\in l} \mathbf{u}_l \right\| \left\| \sum_{i\in V} \mathbf{v}_i \right\| \leq \sqrt{r} \max_{l\in[r]} \|\mathbf{u}_l\| \left\| \sum_{i\in V} \mathbf{v}_i \right\| = \sqrt{r} \left\| \sum_{i\in V} \mathbf{v}_i \right\| . \tag{14}$$

Now, since $\{\mathbf{v}_i\}_{i\in V}$ are feasible solution to r-COL SDP, they form an orthogonal representation of the graph G. By Lemma 14 we have that,

$$\left\| \sum_{i\in V} \mathbf{v}_i \right\| \leq \sqrt{\vartheta(G) \sum_{i\in V} \|\mathbf{v}_i\|^2}. \tag{15}$$

Using constraint (9), and summing over all $i \in V$, we have that,

$$\sum_{i\in V} \|\mathbf{v}_i\|^2 = \sum_{i\in V}\sum_{l\in[r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle .$$

Now using above expression in Eq. (15) it follows that,

$$\left\| \sum_{i\in V} \mathbf{v}_i \right\| \leq \sqrt{\vartheta(G) \sum_{i\in V}\sum_{l\in[r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle}.$$

Now using the above expression in Eq. (14) we obtain that,

$$\sum_{i\in V}\sum_{l\in[r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle \leq \sqrt{r} \left\| \sum_{i\in V} \mathbf{v}_i \right\| \leq \sqrt{r\vartheta(G) \sum_{i\in V}\sum_{l\in[r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle}.$$

Squaring both sides we have that,

$$\sum_{i\in V}\sum_{l\in[r]} \langle \mathbf{v}_i, \mathbf{u}_l \rangle \leq r\vartheta(G).$$

Since the above holds for any feasible $\left\{ \{\mathbf{v}_i\}_{i\in V}, \{\mathbf{u}_l\}_{l\in[r]} \right\}$ of r-COL SDP, it also holds for the optimal solution and we have that,

$$\text{opt}\left(r\text{-COL SDP}\right) \leq r\vartheta(G).$$

■

Next, we recall an important fact, that for a random $G_{n,p}$ graph, the value of Lovasz Theta function $\vartheta(G)$ is bounded above by $\mathcal{O}_p(\sqrt{n})$. Formally we have that,

Fact 16 (Juhász (1982), Lemma 4.2 in Coja-Oghlan (2005)) *For a graph $G \sim G(n,p)$, with parameter regimes of $\frac{(\ln n)^7}{n} \leq p \leq \frac{1}{2}$, we have that with high probability (over the randomness of the input instance), $\vartheta(G) \leq 3\sqrt{\frac{n}{p}}$.*

Proposition 17 *Given an optimal solution to* r-COL SDP*, we define a set* $T_\delta = \left\{ i : \|\mathbf{v}_i^*\|^2 \geq 1 - \delta \right\}$*. If,* $\sum_{i \in S} \|\mathbf{v}_i^*\|^2 \geq (1-\eta)k$*, we have that,* $|T_\delta \cap S| \geq k\left(1 - \frac{\eta}{\delta}\right)$*.*

Proof Let Z be a uniform random variable that takes values in the set $\left\{ \|\mathbf{v}_i^*\|^2 \right\}_{i \in S}$ with probability $1/k$ each. Then using above we have that

$$\mathbb{E}[Z] = \sum_{i \in S} \frac{1}{k} \|\mathbf{v}_i^*\|^2 = \frac{1}{k} \sum_{i \in S} \|\mathbf{v}_i^*\|^2 \geq 1 - \eta.$$

Now from the SDP constraints (9) and (10) we have that,

$$\|\mathbf{v}_i^*\|^2 = \sum_{l \in [r]} \langle \mathbf{v}_i^*, \mathbf{u}_l^* \rangle \leq 1.$$

and therefore, it also follows that $Z \leq 1$. Hence we can apply Markov's inequality on the random variable $1 - Z$.

$$\begin{aligned} \mathbb{P}\left[Z \geq 1 - \delta\right] &= 1 - \mathbb{P}\left[Z \leq 1 - \delta\right] = 1 - \mathbb{P}\left[1 - Z \geq \delta\right] \\ &\geq 1 - \frac{\mathbb{E}[1 - Z]}{\delta} \geq 1 - \frac{\eta}{\delta}. \end{aligned}$$

Since Z is a uniform distribution over $\left\{ \|\mathbf{v}_i^*\|^2 \right\}_{i \in S}$ we have that,

$$|T_\delta \cap S| = \left| \left\{ i \in S : \|\mathbf{v}_i^*\|^2 \geq 1 - \delta \right\} \right| \geq k\left(1 - \frac{\eta}{\delta}\right).$$

■

Proof [Proof of Theorem 9] We recall that in Theorem 12, we showed that our r-COL SDP has opt (r-COL SDP) $\geq k$. Next, we showed in Lemma 15 that,

$$\mathsf{opt}\left(r\text{-COL SDP}\right) \leq r\vartheta(G).$$

For an optimal solution to r-COL SDP we thus have that,

$$k \leq \sum_{i \in V} \sum_{l \in [r]} \langle \mathbf{v}_i^*, \mathbf{u}_l^* \rangle = \sum_{i \in S} \sum_{l \in [r]} \langle \mathbf{v}_i^*, \mathbf{u}_l^* \rangle + \sum_{i \in V \setminus S} \sum_{l \in [r]} \langle \mathbf{v}_i^*, \mathbf{u}_l^* \rangle. \tag{16}$$

We note that the second term is the contribution in the objective from graph induced on $V \setminus S$ which is a random $G_{n-k,p}$ graph in our Definition 1. Consider a r-$\widehat{\text{COL SDP}}$ which is the same as r-COL SDP but only on the graph induced on $V \setminus S$. Therefore, $\left\{ \{\mathbf{v}_i^*\}_{i \in V \setminus S}, \mathbf{u}_1^*, \mathbf{u}_2^*, \ldots, \mathbf{u}_r^* \right\}$ are also a feasible solution to r-$\widehat{\text{COL SDP}}$. Now using Fact 16 we have that,

$$\sum_{i \in V \setminus S} \sum_{l \in [r]} \langle \mathbf{v}_i^*, \mathbf{u}_l^* \rangle \leq \mathsf{opt}\left(r\text{-}\widehat{\text{COL SDP}}\right) \leq 3r\sqrt{\frac{n-k}{p}} \leq 3r\sqrt{\frac{n}{p}}. \tag{17}$$

and using this in Eq. (16) we have that,

$$\sum_{i\in S}\sum_{l\in[r]}\langle \mathbf{v}_i,\mathbf{u}_l\rangle \geq k-3r\sqrt{\frac{n}{p}}.$$

For our regimes of $k \geq \frac{3r}{\varepsilon^2}\sqrt{\frac{n}{p}}$ we thus have that,

$$\sum_{i\in S}\sum_{l\in[r]}\langle \mathbf{v}_i^*,\mathbf{u}_l^*\rangle \geq (1-\varepsilon^2)k.$$

Using our SDP constraint (9) we have that,

$$\sum_{i\in S}\|\mathbf{v}_i^*\|^2=\sum_{i\in S}\sum_{l\in[r]}\langle \mathbf{v}_i^*,\mathbf{u}_l^*\rangle \geq (1-\varepsilon^2)k.$$

Now using Proposition 17 with choice of $\eta=\varepsilon^2$ and $\delta=\varepsilon$ (note that for this choice of δ we have $T=T_\delta$), and we have that

$$|T\cap S| \geq k\left(1-\varepsilon\right).$$

Further using SDP constraint (9) and Eq. (17) we have that,

$$\sum_{i\in V\setminus S}\|\mathbf{v}_i^*\|^2=\sum_{i\in V\setminus S}\sum_{l\in[r]}\langle \mathbf{v}_i^*,\mathbf{u}_l^*\rangle \leq 3r\sqrt{\frac{n}{p}}.$$

Thus we define $T'=\left\{i\in V\setminus S : \|\mathbf{v}_i^*\|^2\geq 1-\varepsilon\right\}$ and we have that,

$$3r\sqrt{\frac{n}{p}}\geq\sum_{i\in V\setminus S}\|\mathbf{v}_i\|^2\geq\left|T'\right|\left(1-\varepsilon\right).$$

Hence, we have for our regimes of that $k\geq\frac{3r}{\varepsilon^2}\sqrt{\frac{n}{p}}$ we have,

$$|T\cap(V\setminus S)|=\left|T'\right|\leq\frac{3r\sqrt{n}}{(1-\varepsilon)\sqrt{p}}\leq\left(\frac{\varepsilon^2}{1-\varepsilon}\right)k.$$

■

Proof [Proof of Corollary 10] We note that adding monotone adversarial edges as in Definition 2, can only increase the number of constraints of r-COL SDP. Let $\widehat{r\text{-COL SDP}}$ be the SDP obtained after adding adversarial edges. Then, since any feasible solution to $\widehat{r\text{-COL SDP}}$ is also a solution to r-COL SDP; we have that,

$$\mathsf{opt}\left(\widehat{r\text{-COL SDP}}\right)\leq\mathsf{opt}\left(r\text{-COL SDP}\right)$$

Now we proceed as in proof of Theorem 9 to obtain similar guarantees. ■

Lemma 18 *[Lemma 43 in Kumar et al. (2022)] Given a graph planted as per Definition 1 for $r = 2$, and given a set S' of size k such that $|S \cap S'| \geq (1-\delta)\,k$ (for $\delta \leq p^2/16$) for the regimes of $k \geq (1152\sqrt{n})/p^3$ and $p \geq 5\sqrt{(\log k)/k}$, there exists a polynomial time deterministic algorithm that recovers the planted set S exactly.*

Corollary 19 *Given a graph $G = (V, E)$ generated according to Definition 1 (for $r = 2$ and d-regular G_S), and regimes of n, k, p satisfying $k \geq (1152\sqrt{n})/p^3$ and $5\sqrt{(\log k)/k} \leq p \leq \frac{1}{2}$, there is a polynomial time algorithm which with high probability (over the randomness of the input instance) recovers the planted set exactly.*

Proof We note that for planted bipartite graphs ($r = 2$), where $\varepsilon \leq \frac{1}{2}$ and $k \geq \frac{6}{\varepsilon^2}\sqrt{\frac{n}{p}}$, Theorem 9 guarantees that $|T \cap S| \geq (1-\varepsilon)\,k$ and that,

$$|T| \leq \left(1 + \frac{\varepsilon^2}{1-\varepsilon}\right) k \leq \left(1 + \frac{\varepsilon^2}{\varepsilon}\right) k = (1+\varepsilon)\,k.$$

Now we arbitrary drop vertices (atmost εk vertices) from the set T and obtain a set S' such that $|S'| = k$. We also note that we have,

$$\left|S' \cap S\right| \geq \left|S' \cap T\right| - \left|T \setminus S'\right| \geq (1-\varepsilon)\,k - \varepsilon k = (1-2\varepsilon)\,k$$

Now by setting $\delta = 2\varepsilon$ we have that $\varepsilon \leq p^2/32$, and we note that,

$$k \geq \max\left\{\frac{1152\sqrt{n}}{p^3}, 192\sqrt{\frac{n}{p^5}}\right\} = \frac{(1152\sqrt{n})}{p^3}.$$

Therefore, using Lemma 18 for the regimes of $k \geq (1152\sqrt{n})/p^3$ and $p \geq 5\sqrt{(\log k)/k}$, we recover the planted bipartite graph exactly. ■

Appendix B. Distinguishing arbitrary r-colorable graph in Random Planted Model

We consider the hypothesis testing formulation of the detection version of the problem as,

- $\mathcal{H}_1 : G \sim G(n, k, p, d)$, a planted r-colorable graph as in Definition 1 against
- $\mathcal{H}_0 \sim G(n, p)$, a random Erdős-Rényi graph

We show that if the planted graph has average degree d such that $d = \Omega(r\sqrt{pn})$, we can distinguish between these two scenarios with high probability (over the randomness of the input). The spectral test is simply to examine the smallest eigenvalue of the adjacency matrix A as,

- If $\lambda_{\min}(A) \leq -\frac{d}{r-1} + (2 + o(1))\sqrt{pn}$, output $\mathcal{H}_1$.
- Otherwise, output $\mathcal{H}_0$.

The test works since for a graph sampled from $G(n,p)$ as in $\mathcal{H}_0$ it is well known (follows from Fact 21) that $\lambda_{\min}(A) \geq -(2+o(1))\sqrt{pn}$ with high probability (over the randomness of the input instance). On the other hand for the graph sampled in $\mathcal{H}_1$, we show next in Lemma 20 that with high probability (over the randomness of the input instance) we have that,

$$\lambda_{\min}(A) \leq -d/(r-1) + (2+o(1))\sqrt{pn}.$$

Thus if we have that $-d/(r-1) + (2+o(1))\sqrt{pn} \leq -(2+o(1))\sqrt{pn}$ which simplifies to $d > (r-1)(4+o(1))\sqrt{pn}$, the test correctly distinguishes with high probability (over the randomness of the input instance).

Lemma 20 *For a graph G generated according to $\mathcal{H}_1$ from Definition 1, with planted subgraph $G[S]$ having average degree $d_{\mathsf{avg}} \geq d$, we have that with high probability (over the randomness of the input instance),*

$$\lambda_{\min}(A) \leq -\frac{d}{r-1} + (2+o(1))\sqrt{pn}.$$

Proof We start by writing the adjacency matrix A of the graph obtained from Definition 1 as,

$$\begin{aligned} A &= A_{S,S} + A_{S,V\setminus S} + A_{V\setminus S,S} + A_{V\setminus S,V\setminus S} \\ &= A_{S,S} + \left(A_{S,V\setminus S} + A_{V\setminus S,S}\right) + p\mathbf{1}_{V\setminus S}\mathbf{1}_{V\setminus S}^T + R \end{aligned} \tag{18}$$

where $A_{U,V}$ is a matrix which is $A[i,j]$ if $i \in U, j \in V$ and 0 otherwise and R is the perturbation matrix that has entries, $R[i,j]$ are zero mean random variables which takes value $-p$ with probability $1-p$ and $-(1-p)$ with probability p when $i \in V \setminus S, j \in V \setminus S$ and are constantly 0 otherwise.

Fact 21 (Spectral Norm of Random Matrices, Füredi and Komlós (1981); Vu (2007)) *For the zero-mean random matrix $R = A - \mathbb{E}[A]$ defined in Eq. (18), its spectral norm satisfies $\|R\| \leq (2+o(1))\sqrt{pn}$ with high probability.*

Claim 22 *For a graph obtained from Definition 1, and $A_{S,S}$ as defined in Eq. (18) we have that,*

$$\lambda_{\min}\left(A_{S,S}\right) \leq -d/(r-1).$$

Proof We will use the following lower bound on chromatic number of graphs due to the works of Hoffman (1970); Hoffman and Howes (1970),

Fact 23 (Hoffman (1970); Hoffman and Howes (1970)) *For a graph G with adjacency matrix A and chromatic number $\chi(G)$ we have,*

$$\chi(G) \geq 1 + \frac{\lambda_{\max}(A)}{-\lambda_{\min}(A)}.$$

Now we note that the graph G' on the k vertices of S and denote the corresponding matrix as A'. Then by definition the graph is r-colorable and using variational characterization of eigenvalues (Theorem 4.2.2 in Horn and Johnson (2013)) we have,

$$\lambda_{\max}(A') = \max_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq 0}} \frac{\mathbf{x}^T A' \mathbf{x}}{\mathbf{x}^T\mathbf{x}} \geq \frac{\left(\mathbf{1}_S/\sqrt{k}\right)^T A' \left(\mathbf{1}_S/\sqrt{k}\right)}{\left(\mathbf{1}_S^T/\sqrt{k}\right)\left(\mathbf{1}_S/\sqrt{k}\right)} = \frac{2\,|E(S,S)|}{k} = \frac{\sum_{i\in S} d(i)}{k} = d_{\mathsf{avg}} \geq d$$

where the first inequality holds with $\mathbf{1}_S/\sqrt{k}$ as a test vector. Now using Fact 23 on G' we have that,

$$r \geq 1 + \frac{d}{-\lambda_{\min}(A')} \text{ which simplifies to } \lambda_{\min}(A') \leq -d/(r-1).$$

Now we let $\mathbf{v}' \in \mathbb{R}^k$ be the unit vector corresponding to the smallest eigenvalue of A'. We consider a vector $\mathbf{v} \in \mathbb{R}^n$ by padding $\mathbf{v}'$ with zeroes corresponding to the entries in $V \setminus S$. Then we have that,

$$\begin{aligned}
\lambda_{\min}(A_{S,S}) = \min_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T A_{S,S}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} &\leq \frac{\mathbf{v}^T A_{S,S}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \frac{\mathbf{v}'^T A_{S,S}\mathbf{v}'}{\mathbf{v}'^T\mathbf{v}'} \\
&= \frac{\mathbf{v}'^T A'\mathbf{v}'}{\mathbf{v}'^T\mathbf{v}'} = \lambda_{\min}(A') \leq -d/(r-1).
\end{aligned}$$

■

In the proof above we note that the vector $\mathbf{v}$ constructed is supported only on the vertices in S. Therefore we have that, $\mathbf{v}^T\left(A_{S,V\setminus S} + A_{V\setminus S,S}\right)\mathbf{v} = 0$ and $\mathbf{v}^T\left(p\mathbf{1}_{V\setminus S}\mathbf{1}_{V\setminus S}^T\right)\mathbf{v} = 0$, and $\mathbf{v}$ is a unit vector. Further using Fact 21 we have that with high probability (over the randomness of the input instance),

$$\begin{aligned}
\lambda_{\min}(A) = \min_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} &\leq \frac{\mathbf{v}^T A\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \mathbf{v}^T A\mathbf{v} \\
&= \mathbf{v}^T\left(A_{S,S} + \left(A_{S,V\setminus S} + A_{V\setminus S,S}\right) + p\mathbf{1}_{V\setminus S}\mathbf{1}_{V\setminus S}^T + R\right)\mathbf{v} \\
&\leq \mathbf{v}'^T A'\mathbf{v}' + \mathbf{v}^T\left(A_{S,V\setminus S} + A_{V\setminus S,S}\right)\mathbf{v} + \mathbf{v}^T\left(p\mathbf{1}_{V\setminus S}\mathbf{1}_{V\setminus S}^T\right)\mathbf{v} + \mathbf{v}^T R\mathbf{v} \\
&\leq -\frac{d}{(r-1)} + (2+o(1))\sqrt{pn}.
\end{aligned}$$

■