JMLR: Workshop and Conference Proceedings vol 40:1–30, 2015

Computational Lower Bounds for Community Detection on Random Graphs

Bruce Hajek B-HAJEK@ILLINOIS.EDU
Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL

Yihong Wu YIHONGWU@ILLINOIS.EDU
Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL

Jiaming Xu JIAMINGX@WHARTON.UPENN.EDU
Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA,

Abstract

This paper studies the problem of detecting the presence of a small dense community planted in a large Erdős-Rényi random graph $\mathcal{G}(N, q)$, where the edge probability within the community exceeds q by a constant factor. Assuming the hardness of the planted clique detection problem, we show that the computational complexity of detecting the community exhibits the following phase transition phenomenon: As the graph size N grows and the graph becomes sparser according to $q = N^{-\alpha}$, there exists a critical value of $\alpha = \frac{2}{3}$, below which there exists a computationally intensive procedure that can detect far smaller communities than any computationally efficient procedure, and above which a linear-time procedure is statistically optimal. The results also lead to the average-case hardness results for recovering the dense community and approximating the densest K-subgraph.

1. Introduction

Networks often exhibit community structure with many edges joining the vertices of the same community and relatively few edges joining vertices of different communities. Detecting communities in networks has received a large amount of attention and has found numerous applications in social and biological sciences, etc (see, e.g., the exposition Fortunato (2010) and the references therein). While most previous work focuses on identifying the vertices in the communities, this paper studies the more basic problem of detecting the presence of a small community in a large random graph, proposed recently in Arias-Castro and Verzelen (2014). This problem has practical applications including detecting new events and monitoring clusters, and is also of theoretical interest for understanding the statistical and algorithmic limits of community detection Chen and Xu (2014).

Inspired by the model in Arias-Castro and Verzelen (2014), we formulate this community detection problem as a planted dense subgraph detection (PDS) problem. Specifically, let $\mathcal{G}(N, q)$ denote the Erdős-Rényi random graph with N vertices, where each pair of vertices is connected independently with probability q. Let $\mathcal{G}(N, K, p, q)$ denote the planted dense subgraph model with N vertices where: (1) each vertex is included in the random set S independently with probability $\frac{K}{N}$; (2) for any two vertices, they are connected independently with probability p if both of them are in S and with probability q otherwise, where $p > q$. In this case, the vertices in S form a community with higher connectivity than elsewhere. The planted dense subgraph here has a random size

with mean K, instead of a deterministic size K as assumed in Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013).

Definition 1 *The planted dense subgraph detection problem with parameters (N, K, p, q), henceforth denoted by $\mathsf{PDS}(N, K, p, q)$, refers to the problem of distinguishing hypotheses:*

$$H_0: \quad G \sim \mathcal{G}(N, q) \triangleq \mathbb{P}_0, \qquad H_1: \quad G \sim \mathcal{G}(N, K, p, q) \triangleq \mathbb{P}_1.$$

The statistical difficulty of the problem depends on the parameters (N, K, p, q). Intuitively, if the expected dense subgraph size K decreases, or if the edge probabilities p and q both decrease by the same factor, or if p decreases for q fixed, the distributions under the null and alternative hypotheses become less distinguishable. Recent results in Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013) obtained necessary and sufficient conditions for detecting planted dense subgraphs under certain assumptions of the parameters. However, it remains unclear whether the statistical fundamental limit can always be achieved by efficient procedures. In fact, it has been shown in Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013) that many popular low-complexity tests, such as total degree test, maximal degree test, dense subgraph test, as well as tests based on certain convex relaxations, can be highly suboptimal. This observation prompts us to investigate the computational limits for the PDS problem, i.e., what is the sharp condition on (N, K, p, q) under which the problem admits a computationally efficient test with vanishing error probability, and conversely, without which no algorithm can detect the planted dense subgraph reliably in polynomial time. To this end, we focus on a particular case where the community is denser by a constant factor than the rest of the graph, i.e., $p = cq$ for some constant $c > 1$. Adopting the standard reduction approach in complexity theory, we show that the PDS problem in some parameter regime is at least as hard as the planted clique problem in some parameter regime, which is conjectured to be computationally intractable. Let $\mathcal{G}(n, k, \gamma)$ denote the planted clique model in which we add edges to k vertices uniformly chosen from $\mathcal{G}(n, \gamma)$ to form a clique.

Definition 2 *The PC detection problem with parameters (n, k, γ), denoted by $\mathsf{PC}(n, k, \gamma)$ henceforth, refers to the problem of distinguishing hypotheses:*

$$H_0^{\mathrm{C}}: \quad G \sim \mathcal{G}(n, \gamma), \qquad H_1^{\mathrm{C}}: \quad G \sim \mathcal{G}(n, k, \gamma).$$

The problem of finding the planted clique has been extensively studied for $\gamma = \frac{1}{2}$ and the state-of-the-art polynomial-time algorithms Alon et al. (1998); Feige and Krauthgamer (2000); McSherry (2001); Feige and Ron (2010); Dekel et al. (2010); Ames and Vavasis (2011); Deshpande and Montanari (2012) only work for $k = \Omega(\sqrt{n})$. There is no known polynomial-time solver for the PC problem for $k = o(\sqrt{n})$ and any constant $\gamma > 0$. It is conjectured Jerrum (1992); Hazan and Krauthgamer (2011); Juels and Peinado (2000); Alon et al. (2007); Feldman et al. (2013) that the PC problem cannot be solved in polynomial time for $k = o(\sqrt{n})$ with $\gamma = \frac{1}{2}$, which we refer to as the PC Hypothesis.

Hypothesis 1 *Fix some constant $0 < \gamma \leq \frac{1}{2}$. For any sequence of randomized polynomial-time tests $\{\psi_{n,k_n}\}$ such that $\limsup_{n\to\infty} \frac{\log k_n}{\log n} < 1/2$,*

$$\liminf_{n\to\infty} \mathbb{P}_{H_0^{\mathrm{C}}}\{\psi_{n,k}(G) = 1\} + \mathbb{P}_{H_1^{\mathrm{C}}}\{\psi_{n,k}(G) = 0\} \geq 1.$$

The PC Hypothesis with $\gamma = \frac{1}{2}$ is similar to (Ma and Wu, 2015, Hypothesis 1) and (Berthet and Rigollet, 2013, Hypothesis $\mathbf{B_{PC}}$). Our computational lower bounds require that the PC Hypothesis holds for any positive constant γ. An even stronger assumption that PC Hypothesis holds for $\gamma = 2^{-\log^{0.99} n}$ has been used in (Applebaum et al., 2010, Theorem 10.3) for public-key cryptography. Furthermore, (Feldman et al., 2013, Corollary 5.8) shows that under a statistical query model, any statistical algorithm requires at least $n^{\Omega(\frac{\log n}{\log(1/\gamma)})}$ queries for detecting the planted bi-clique in an Erdős-Rényi random bipartite graph with edge probability γ.

1.1. Main Results

We consider the $\mathsf{PDS}(N, K, p, q)$ problem in the following asymptotic regime:

$$p = cq = \Theta(N^{-\alpha}),\ K = \Theta(N^{\beta}), \quad N \to \infty, \tag{1}$$

where $c > 1$ is a fixed constant, $\alpha \in [0, 2]$ governs the sparsity of the graph,[1] and $\beta \in [0, 1]$ captures the size of the dense subgraph. Clearly the detection problem becomes more difficult if either α increases or β decreases. Assuming the PC Hypothesis holds for any positive constant γ, we show that the parameter space of (α, β) is partitioned into three regimes as depicted in Fig. 1:

- **The Simple Regime:** $\beta > \frac{1}{2} + \frac{\alpha}{4}$. The dense subgraph can be detected in linear time with high probability by thresholding the total number of edges.
- **The Hard Regime:** $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$. Reliable detection can be achieved by thresholding the maximum number of edges among all subgraphs of size K; however, no polynomial-time solver exists in this regime.
- **The Impossible Regime:** $\beta < \min\{\alpha, \frac{1}{2} + \frac{\alpha}{4}\}$. No test can detect the planted subgraph regardless of the computational complexity.

The computational hardness of the PDS problem exhibits a phase transition at the critical value $\alpha = 2/3$: For *moderately sparse* graphs with $\alpha < 2/3$, there exists a combinatorial algorithm that can detect far smaller communities than any efficient procedures; For *highly sparse* graphs with $\alpha > 2/3$, optimal detection is achieved in linear time based on the total number of edges. Equivalently, attaining the statistical detection limit is computationally tractable only in the large-community regime ($\beta > 2/3$). Therefore, surprisingly, the linear-time test based on the total number of edges is always statistically optimal among all computationally efficient procedures in the sense that no polynomial-time algorithm can reliably detect the community when $\beta < \frac{1}{2} + \frac{\alpha}{4}$. It should be noted that Fig. 1 only captures the leading polynomial term according to the parametrization (1); at the boundary $\beta = \alpha/4 + 1/2$, it is plausible that one needs to go beyond simple edge counting in order to achieve reliable detection. This is analogous to the planted clique problem where the maximal degree test succeeds if the clique size satisfies $k = \Omega(\sqrt{n \log n})$ Kučera (1995) and the more sophisticated spectral method succeeds if $k = \Omega(\sqrt{n})$ Alon et al. (1998).

The above hardness result should be contrasted with the recent study of community detection on the stochastic block model, where the community size scales linearly with the network size. When the edge density scales as $\Theta(\frac{1}{N})$ Mossel et al. (2012, 2013); Massoulié (2013) (resp. $\Theta(\frac{\log N}{N})$

1. The case of $\alpha > 2$ is not interesting since detection is impossible even if the planted subgraph is the entire graph ($K = N$).

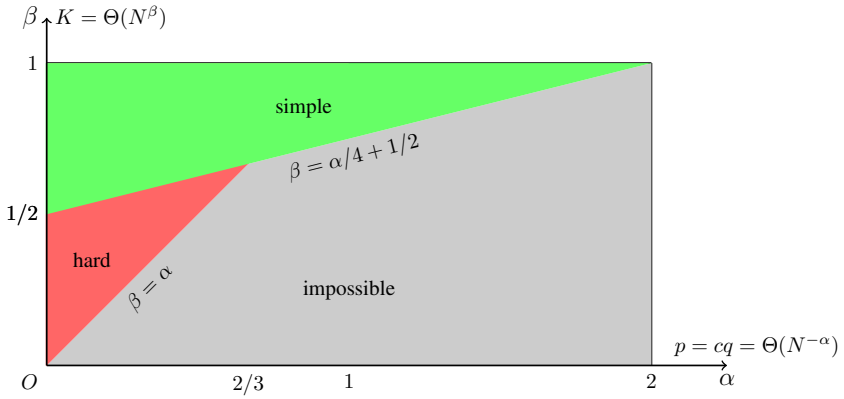

Figure 1: The simple (green), hard (red), impossible (gray) regimes for detecting the planted dense subgraph.

Abbe et al. (2014); Mossel et al. (2014); Hajek et al. (2014)), the statistically optimal threshold for partial (resp. exact) recovery can be attained in polynomial time up to the sharp constants. In comparison, this paper focuses on the regime when the community size grows *sublinearly* as N^β and the edge density decays more slowly as $N^{-\alpha}$. It turns out that in this case even achieving the optimal exponent is computationally as demanding as solving the planted clique problem.

Our computational lower bound for the PDS problem also implies the average-case hardness of approximating the planted dense subgraph or the densest K-subgraph of the random graph ensemble $\mathcal{G}(N, K, p, q)$, complementing the worst-case inapproximability result in Alon et al. (2011), which is based on the planted clique hardness as well. In particular, we show that no polynomial-time algorithm can approximate the planted dense subgraph or the densest K-subgraph within any constant factor in the regime of $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$, which provides a partial answer to the conjecture made in (Chen and Xu, 2014, Conjecture 2.6) and the open problem raised in (Alon et al., 2011, Section 4) (see Section 4.1). Our approach and results can be extended to the bipartite graph case (see Section 4.3) and shed light on the computational limits of the PDS problem with a fixed planted dense subgraph size studied in Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013) (see Section 4.2).

1.2. Connections to the Literature

This work is inspired by an emerging line of research (see, e.g., Kolar et al. (2011); Balakrishnan et al. (2011); Berthet and Rigollet (2013); Chandrasekaran and Jordan (2013); Ma and Wu (2015); Chen and Xu (2014); Xu et al. (2014)) which examines high-dimensional inference problems from both the statistical and computational perspectives. Our computational lower bounds follow from a randomized polynomial-time reduction scheme which approximately reduces the PC problem to the PDS problem of appropriately chosen parameters. Below we discuss the connections to previous results and highlight the main technical contributions of this paper.

PC Hypothesis Various hardness results in the theoretical computer science literature have been established based on the PC Hypothesis with $\gamma = \frac{1}{2}$, e.g. cryptographic applications Juels and Peinado (2000), approximating Nash equilibrium Hazan and Krauthgamer (2011), testing k-wise

independence Alon et al. (2007), etc. More recently, the PC Hypothesis with $\gamma = \frac{1}{2}$ has been used to investigate the penalty incurred by complexity constraints on certain high-dimensional statistical inference problems, such as detecting sparse principal components Berthet and Rigollet (2013) and noisy biclustering (submatrix detection) Ma and Wu (2015). Compared with most previous works, our computational lower bounds rely on the stronger assumption that the PC Hypothesis holds for any positive constant γ. An even stronger assumption that PC Hypothesis holds for $\gamma = 2^{-\log^{0.99} n}$ has been used in Applebaum et al. (2010) for public-key cryptography. It is an interesting open problem to prove that PC Hypothesis for a fixed $\gamma \in (0, \frac{1}{2})$ follows from that for $\gamma = \frac{1}{2}$.

Reduction from the PC Problem Most previous work Hazan and Krauthgamer (2011); Alon et al. (2007, 2011); Applebaum et al. (2010) in the theoretical computer science literature uses the reduction from the PC problem to generate computationally hard instances of problems and establish *worst-case* hardness results; the underlying distributions of the instances could be arbitrary. The idea of proving hardness of a hypothesis testing problem by means of approximate reduction from the planted clique problem such that the reduced instance is close to the target hypothesis in total variation originates from the seminal work by Berthet and Rigollet (2013) and the subsequent paper by Ma and Wu (2015). The main distinction between these results and the present paper is that Berthet and Rigollet (2013) studied a composite-versus-composite testing problem and Ma and Wu (2015) studied a simple-versus-composite testing problem, both in the minimax sense, as opposed to the simple-versus-simple hypothesis considered in this paper. For composite hypothesis, a reduction scheme works as long as the distribution of the reduced instance is close to some mixture under the hypothesis. This freedom is absent in constructing reduction for simple hypothesis, which renders the reduction scheme as well as the corresponding calculation of total variation in the present paper considerably more difficult. For example, Ma and Wu (2015) studied testing a matrix θ is either zero or block-sparse in Gaussian noise. Their reduction maps the alternative distribution of the planted clique to a Gaussian mixtures with respect to some prior over block-sparse matrices. However, their hardness result does not carry over if the prior is predetermined, say, uniform. In contrast, for community detection problems, the goal is to establish the hardness of testing two simple hypothesis, namely, $\mathcal{G}(N, q)$ versus $\mathcal{G}(N, K, p, q)$. Thus the underlying distributions of the problem instances generated from the reduction must be close to the desired distributions in total variation under both the null and alternative hypotheses. To this end, we start with a small dense graph generated from $\mathcal{G}(n, \gamma)$ under H_0 and $\mathcal{G}(n, k, \gamma)$ under H_1, and arrive at a larger sparse graph whose distribution is exactly $\mathcal{G}(N, q)$ under H_0 and approximately equal to $\mathcal{G}(N, K, p, q)$ under H_1. Notice that simply sparsifying the PC problem does not capture the desired tradeoff between the graph sparsity and the cluster size. Our reduction scheme differs from those used in Berthet and Rigollet (2013); Ma and Wu (2015) which start with a large dense graph. Similar to ours, the reduction scheme in Alon et al. (2011) also enlarges and sparsifies the graph by taking its subset power; but the distributions of the resulting random graphs are rather complicated and not close to the Erdős-Rényi type. The techniques of bounding the total variation distance also differs substantially from those used in the previous work Berthet and Rigollet (2013); Ma and Wu (2015). Notably, the use of the theory of associated random variables Dubhashi and Ranjan (1998) is a major new ingredient in the proof.

Inapproximability of the DKS Problem The densest K-subgraph (DKS) problem refers to finding the subgraph of K vertices with the maximal number of edges. In view of the NP-hardness of the DKS problem which follows from the NP-hardness of MAXCLIQUE, it is of interest to con-

sider an η-factor approximation algorithm, which outputs a subgraph with K vertices containing at least a $\frac{1}{\eta}$-fraction of the number of edges in the densest K-subgraph. Proving the NP-hardness of $(1+\epsilon)$-approximation for DKS for any fixed $\epsilon > 0$ is a longstanding open problem. See Alon et al. (2011) for a comprehensive discussion. Assuming the PC Hypothesis holds with $\gamma = \frac{1}{2}$, Alon et al. (2011) shows that the DKS problem is hard to approximate within any constant factor even if the densest K-subgraph is a clique of size $K = N^\beta$ for any $\beta < 1$, where N denotes the total number of vertices. This worst-case inapproximability result is in stark contrast to the average-case behavior in the planted dense subgraph model $G(N, K, p, q)$ under the scaling (1), where it is known Chen and Xu (2014); Ames (2013) that the planted dense subgraph can be exactly recovered in polynomial time if $\beta > \frac{1}{2} + \frac{\alpha}{2}$ (see the simple region in Fig. 2 below), implying that the densest K-subgraph can be approximated within a factor of $1+\epsilon$ in polynomial time for any $\epsilon > 0$. On the other hand, our computational lower bound for $\mathsf{PDS}(N, K, p, q)$ shows that any constant-factor approximation of the densest K-subgraph has high average-case hardness if $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$ (see Section 4.1).

Variants of PDS Model Three versions of the PDS model were considered in (Bhaskara et al., 2010, Section 3). Under all three the graph under the null hypothesis is the Erdős-Rényi graph. The versions of the alternative hypothesis, in order of increasing difficulty of detection, are: (1) The *random planted* model, such that the graph under the alternative hypothesis is obtained by generating an Erdős-Rényi graph, selecting K nodes arbitrarily, and then resampling the edges among the K nodes with a higher probability to form a denser Erdős-Rényi subgraph. This is somewhat more difficult to detect than the model of Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013), for which the choice of which K nodes are in the planted dense subgraph is made before any edges of the graph are independently, randomly generated. (2) The *dense in random* model, such that both the nodes and edges of the planted dense K-subgraph are arbitrary; (3) The *dense versus random* model, such that the entire graph under the alternative hypothesis could be an arbitrary graph containing a dense K-subgraph. Our PDS model is closely related to the first of these three versions, the key difference being that for our model the size of the planted dense subgraph is binomially distributed with mean K (see Section 4.2). Thus, our hardness result is for the easiest type of detection problem. A bipartite graph variant of the PDS model is used in (Arora et al., 2010, p. 10) for financial applications where the total number of edges is the same under both the null and alternative hypothesis. A hypergraph variant of the PDS problem is used in Applebaum et al. (2010) for cryptographic applications.

1.3. Notations

For any set S, let $|S|$ denote its cardinality. Let $s_1^n = \{s_1, \ldots, s_n\}$. For any positive integer N, let $[N] = \{1, \ldots, N\}$. For $a, b \in \mathbb{R}$, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. We use standard big O notations, e.g., for any sequences $\{a_n\}$ and $\{b_n\}$, $a_n = \Theta(b_n)$ if there is an absolute constant $C > 0$ such that $1/C \le a_n/b_n \le C$. Let $\mathrm{Bern}(p)$ denote the Bernoulli distribution with mean p and $\mathrm{Binom}(N, p)$ denote the binomial distribution with N trials and success probability p. For random variables X, Y, we write $X \perp\!\!\!\perp Y$ if X is independent with Y. For probability measures $\mathbb{P}$ and $\mathbb{Q}$, let $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\int |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}|$ denote the total variation distance and $\chi^2(\mathbb{P}\|\mathbb{Q}) = \int \frac{(\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q})^2}{\mathrm{d}\mathbb{Q}}$ the χ^2-divergence. The distribution of a random variable X is denoted by P_X. We write $X \sim \mathbb{P}$ if $P_X = \mathbb{P}$. All logarithms are natural unless the base is explicitly specified.

2. Statistical Limits

This section determines the statistical limit for the $\mathsf{PDS}(N, K, p, q)$ problem with $p = cq$ for a fixed constant $c > 1$. For a given pair (N, K), one can ask the question: What is the smallest density q such that it is possible to reliably detect the planted dense subgraph? When the subgraph size K is *deterministic*, this question has been thoroughly investigated by Arias-Castro and Verzelen Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013) for general (N, K, p, q) and the statistical limit with sharp constants has obtained in certain asymptotic regime. Their analysis treats the dense regime $\log(1 \vee (Kq)^{-1}) = o(\log \frac{N}{K})$ Arias-Castro and Verzelen (2014) and sparse regime $\log \frac{N}{K} = O(\log(1 \vee (Kq)^{-1}))$ Verzelen and Arias-Castro (2013) separately. Here as we focus on the special case of $p = cq$ and are only interested in characterizations within absolute constants, we provide a simple non-asymptotic analysis which treats the dense and sparse regimes in a unified manner. Our results demonstrate that the PDS problem in Definition 1 has the same statistical detection limit as the PDS problem with a deterministic size K studied in Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013).

2.1. Lower Bound

By the definition of the total variation distance, the optimal testing error probability is determined by the total variation distance between the distributions under the null and the alternative hypotheses:

$$\min_{\phi:\{0,1\}^{N(N-1)/2}\to\{0,1\}} (\mathbb{P}_0\{\phi(G) = 1\} + \mathbb{P}_1\{\phi(G) = 0\}) = 1 - d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_1).$$

The following result (proved in Section A.1) shows that if $q = O(\frac{1}{K} \log \frac{eN}{K} \wedge \frac{N^2}{K^4})$, then there exists no test which can detect the planted subgraph reliably.

Proposition 3 *Suppose $p = cq$ for some constant $c > 1$. There exists a function $h : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $h(0+) = 0$ such that the following holds: For any $1 \le K \le N$, $C > 0$ and $q \le C(\frac{1}{K} \log \frac{eN}{K} \wedge \frac{N^2}{K^4})$,*

$$d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_1) \le h(Cc^2) + \exp(-K/8). \tag{2}$$

2.2. Upper Bound

Let A denote the adjacency matrix of the graph G. The detection limit can be achieved by the linear test statistic and scan test statistic proposed in Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013):

$$T_{\mathrm{lin}} \triangleq \sum_{i<j} A_{ij}, \quad T_{\mathrm{scan}} \triangleq \max_{S':|S'|=K} \sum_{i,j\in S':i<j} A_{ij}, \tag{3}$$

which correspond to the total number of edges in the whole graph and the densest K-subgraph, respectively. Interestingly, the exact counterparts of these tests have been proposed and shown to be minimax optimal for detecting submatrices in Gaussian noise Butucea and Ingster (2013); Kolar et al. (2011); Ma and Wu (2015). The following lemma bounds the error probabilities of the linear and scan test.

Proposition 4 *Suppose $p = cq$ for a constant $c > 1$. For the linear test statistic, set $\tau_1 = \binom{N}{2}q + \binom{K}{2}(p-q)/2$. For the scan test statistic, set $\tau_2 = \binom{K}{2}(p+q)/2$. Then there exists a constant C which only depends on c such that*

$$\mathbb{P}_0[T_{\mathrm{lin}} > \tau_1] + \mathbb{P}_1[T_{\mathrm{lin}} \le \tau_1] \le 2\exp\left(-C\frac{K^4 q}{N^2}\right) + \exp\left(-\frac{K}{200}\right)$$
$$\mathbb{P}_0[T_{\mathrm{scan}} > \tau_2] + \mathbb{P}_1[T_{\mathrm{scan}} \le \tau_2] \le 2\exp\left(K\log\frac{Ne}{K} - CK^2 q\right) + \exp\left(-\frac{K}{200}\right).$$

To illustrate the implications of the above lower and upper bounds, consider the $\mathsf{PDS}(N, K, p, q)$ problem with the parametrization $p = cq$, $q = N^{-\alpha}$ and $K = N^{\beta}$ for $\alpha > 0$ and $\beta \in (0, 1)$ and $c > 1$. In this asymptotic regime, the fundamental detection limit is characterized by the following function

$$\beta^*(\alpha) \triangleq \alpha \wedge \left(\frac{1}{2} + \frac{\alpha}{4}\right), \tag{4}$$

which gives the statistical boundary in Fig. 1. Indeed, if $\beta < \beta^*(\alpha)$, as a consequence of Proposition 3, $\mathbb{P}_0\{\phi(G) = 1\} + \mathbb{P}_1\{\phi(G) = 0\} \to 1$ for any sequence of tests. Conversely, if $\beta > \beta^*(\alpha)$, then Proposition 4 implies that the test $\phi(G) = \mathbf{1}_{\{T_{\mathrm{lin}} > \tau_1 \text{ or } T_{\mathrm{scan}} > \tau_2\}}$ achieves vanishing Type-I+II error probabilities. More precisely, the linear test succeeds in the regime $\beta > \frac{1}{2} + \frac{\alpha}{4}$, while the scan test succeeds in the regime $\beta > \alpha$.

Note that T_{lin} can be computed in linear time. However, computing T_{scan} amounts to enumerating all subsets of $[N]$ of cardinality K, which can be computationally intensive. Therefore it is unclear whether there exists a polynomial-time solver in the regime $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$. Assuming the PC Hypothesis, this question is resolved in the negative in the next section.

3. Computational Lower Bounds

In this section, we establish the computational lower bounds for the PDS problem assuming the intractability of the planted clique problem. We show that the PDS problem can be approximately reduced from the PC problem of appropriately chosen parameters in randomized polynomial time. Based on this reduction scheme, we establish a formal connection between the PC problem and the PDS problem in Proposition 5, and the desired computational lower bounds follow from Theorem 7.

We aim to reduce the $\mathsf{PC}(n, k, \gamma)$ problem to the $\mathsf{PDS}(N, K, cq, q)$ problem. For simplicity, we focus on the case of $c = 2$; the general case follows similarly with a change in some numerical constants that come up in the proof. We are given an adjacency matrix $A \in \{0,1\}^{n\times n}$, or equivalently, a graph G, and with the help of additional randomness, will map it to an adjacency matrix $\widetilde{A} \in \{0,1\}^{N\times N}$, or equivalently, a graph $\widetilde{G}$ such that the hypothesis H_0^{C} (resp. H_1^{C}) in Definition 2 is mapped to H_0 exactly (resp. H_1 approximately) in Definition 1. In other words, if A is drawn from $\mathcal{G}(n, \gamma)$, then $\widetilde{A}$ is distributed according to $\mathbb{P}_0$; If A is drawn from $\mathcal{G}(n, k, 1, \gamma)$, then the distribution of $\widetilde{A}$ is close in total variation to $\mathbb{P}_1$.

Our reduction scheme works as follows. Each vertex in $\widetilde{G}$ is randomly assigned a parent vertex in G, with the choice of parent being made independently for different vertices in $\widetilde{G}$, and uniformly over the set $[n]$ of vertices in G. Let V_s denote the set of vertices in $\widetilde{G}$ with parent $s \in [n]$ and let $\ell_s = |V_s|$. Then the set of children nodes $\{V_s : s \in [n]\}$ form a random partition of $[N]$. For any

$1 \le s \le t \le n$, the number of edges, $E(V_s, V_t)$, from vertices in V_s to vertices in V_t in $\widetilde{G}$ will be selected randomly with a conditional probability distribution specified below. Given $E(V_s, V_t)$, the particular set of edges with cardinality $E(V_s, V_t)$ is chosen uniformly at random.

It remains to specify, for $1 \le s \le t \le n$, the conditional distribution of $E(s,t)$ given l_s, l_t, and $A_{s,t}$. Ideally, conditioned on ℓ_s and ℓ_t, we want to construct a Markov kernel from $A_{s,t}$ to $E(s,t)$ which maps $\mathrm{Bern}(1)$ to the desired edge distribution $\mathrm{Binom}(\ell_s\ell_t, p)$, and $\mathrm{Bern}(1/2)$ to $\mathrm{Binom}(\ell_s\ell_t, q)$, depending on whether both s and t are in the clique or not, respectively. Such a kernel, unfortunately, provably does not exist. Nonetheless, this objective can be accomplished approximately in terms of the total variation. For $s = t \in [n]$, let $E(V_s, V_t) \sim \mathrm{Binom}(\binom{\ell_t}{2}, q)$. For $1 \le s < t \le n$, denote $P_{\ell_s\ell_t} \triangleq \mathrm{Binom}(\ell_s\ell_t, p)$ and $Q_{\ell_s\ell_t} \triangleq \mathrm{Binom}(\ell_s\ell_t, q)$. Fix $0 < \gamma \le \frac{1}{2}$ and put $m_0 \triangleq \lfloor \log_2(1/\gamma) \rfloor$. Define

$$P'_{\ell_s\ell_t}(m) = \begin{cases} P_{\ell_s\ell_t}(m) + a_{\ell_s\ell_t} & \text{for } m = 0, \\ P_{\ell_s\ell_t}(m) & \text{for } 1 \le m \le m_0, \\ \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) & \text{for } m_0 < m \le \ell_s\ell_t. \end{cases}$$

where $a_{\ell_s\ell_t} = \sum_{m_0 < m \le \ell_s\ell_t} [P_{\ell_s\ell_t}(m) - \frac{1}{\gamma} Q_{\ell_s\ell_t}(m)]$. Let $Q'_{\ell_s\ell_t} = \frac{1}{1-\gamma}(Q_{\ell_s\ell_t} - \gamma P'_{\ell_s\ell_t})$. As we show later, $Q'_{\ell_s\ell_t}$ and $P'_{\ell_s\ell_t}$ are well-defined probability distributions as long as $\ell_s, \ell_t \le 2\ell$ and $16q\ell^2 \le 1$, where $\ell = N/n$. Then, for $1 \le s < t \le n$, let the conditional distribution of $E(V_s, V_t)$ given ℓ_s, ℓ_t, and $A_{s,t}$ be given by

$$E(V_s, V_t) \sim \begin{cases} P'_{\ell_s\ell_t} & \text{if } A_{st} = 1, \ell_s, \ell_t \le 2\ell \\ Q'_{\ell_s\ell_t} & \text{if } A_{st} = 0, \ell_s, \ell_t \le 2\ell \\ Q_{\ell_s\ell_t} & \text{if } \max\{\ell_s, \ell_t\} > 2\ell. \end{cases} \tag{5}$$

The next proposition (proved in Section A.3) shows that the randomized reduction defined above maps $\mathcal{G}(n, \gamma)$ into $\mathcal{G}(N, q)$ under the null hypothesis and $\mathcal{G}(n, k, \gamma)$ approximately into $\mathcal{G}(N, K, p, q)$ under the alternative hypothesis, respectively. The intuition behind the reduction scheme is as follows: By construction, $(1-\gamma)Q'_{\ell_s\ell_t} + \gamma P'_{\ell_s\ell_t} = Q_{\ell_s\ell_t} = \mathrm{Binom}(\ell_s\ell_t, q)$ and therefore the null distribution of the PC problem is exactly matched to that of the PDS problem, i.e., $P_{\widetilde{G}|H_0^C} = \mathbb{P}_0$. The core of the proof lies in establishing that the alternative distributions are approximately matched. The key observation is that $P'_{\ell_s\ell_t}$ is close to $P_{\ell_s\ell_t} = \mathrm{Binom}(\ell_s\ell_t, p)$ and thus for nodes with distinct parents $s \ne t$ in the planted clique, the number of edges $E(V_s, V_t)$ is approximately distributed as the desired $\mathrm{Binom}(\ell_s\ell_t, p)$; for nodes with the same parent s in the planted clique, even though $E(V_s, V_s)$ is distributed as $\mathrm{Binom}(\binom{\ell_s}{2}, q)$ which is not sufficiently close to the desired $\mathrm{Binom}(\binom{\ell_s}{2}, p)$, after averaging over the random partition $\{V_s\}$, the total variation distance becomes negligible.

Proposition 5 *Let $\ell, n \in \mathbb{N}$, $k \in [n]$ and $\gamma \in (0, \frac{1}{2}]$. Let $N = \ell n$, $K = k\ell$, $p = 2q$ and $m_0 = \lfloor \log_2(1/\gamma) \rfloor$. Assume that $16q\ell^2 \le 1$ and $k \ge 6e\ell$. If $G \sim \mathcal{G}(n, \gamma)$, then $\widetilde{G} \sim \mathcal{G}(N, q)$, i.e., $P_{\widetilde{G}|H_0^C} = \mathbb{P}_0$. If $G \sim \mathcal{G}(n, k, 1, \gamma)$, then*

$$d_{\mathrm{TV}}\left(P_{\widetilde{G}|H_1^C}, \mathbb{P}_1\right) \le e^{-\frac{K}{12}} + 1.5ke^{-\frac{\ell}{18}} + 2k^2(8q\ell^2)^{m_0+1} + 0.5\sqrt{e^{72e^2q\ell^2} - 1} + \sqrt{0.5k}e^{-\frac{\ell}{36}}. \tag{6}$$

An immediate consequence of Proposition 5 is the following result (proved in Section A.4) showing that any PDS solver induces a solver for a corresponding instance of the PC problem.

Proposition 6 *Let the assumption of Proposition 5 hold. Suppose* $\phi : \{0,1\}^{\binom{N}{2}} \to \{0,1\}$ *is a test for* $\mathsf{PDS}(N, K, 2q, q)$ *with Type-I+II error probability* η*. Then* $G \mapsto \phi(\widetilde{G})$ *is a test for the* $\mathsf{PC}(n, k, \gamma)$ *whose Type-I+II error probability is upper bounded by* $\eta + \xi$ *with* ξ *given by the right-hand side of (6).*

The following theorem establishes the computational limit of the PDS problem as shown in Fig. 1.

Theorem 7 *Assume Hypothesis 1 holds for a fixed* $0 < \gamma \leq 1/2$*. Let* $m_0 = \lfloor \log_2(1/\gamma) \rfloor$*. Let* $\alpha > 0$ *and* $0 < \beta < 1$ *be such that*

$$\alpha < \beta < \frac{1}{2} + \frac{m_0\alpha + 4}{4m_0\alpha + 4}\alpha - \frac{2}{m_0\alpha}. \tag{7}$$

Then there exists a sequence $\{(N_\ell, K_\ell, q_\ell)\}_{\ell \in \mathbb{N}}$ *satisfying* $\lim_{\ell\to\infty} \frac{\log(1/q_\ell)}{\log N_\ell} = \alpha$ *and* $\lim_{\ell\to\infty} \frac{\log K_\ell}{\log N_\ell} = \beta$ *such that for any sequence of randomized polynomial-time tests* $\phi_\ell : \{0,1\}^{\binom{N_\ell}{2}} \to \{0,1\}$ *for the* $\mathsf{PDS}(N_\ell, K_\ell, 2q_\ell, q_\ell)$ *problem, the Type-I+II error probability is lower bounded by*

$$\liminf_{\ell\to\infty} \mathbb{P}_0\{\phi_\ell(G') = 1\} + \mathbb{P}_1\{\phi_\ell(G') = 0\} \geq 1,$$

where $G' \sim \mathcal{G}(N, q)$ *under* H_0 *and* $G' \sim \mathcal{G}(N, K, p, q)$ *under* H_1*. Consequently, if Hypothesis 1 holds for all* $0 < \gamma \leq 1/2$*, then the above holds for all* $\alpha > 0$ *and* $0 < \beta < 1$ *such that*

$$\alpha < \beta < \beta^\sharp(\alpha) \triangleq \frac{1}{2} + \frac{\alpha}{4}. \tag{8}$$

Consider the asymptotic regime given by (1). The function $\beta^\sharp$ in (8) gives the computational barrier for the $\mathsf{PDS}(N, K, p, q)$ problem (see Fig. 1). Compared to the statistical limit β^* given in (4), we note that $\beta^*(\alpha) < \beta^\sharp(\alpha)$ if and only if $\alpha < \frac{2}{3}$, in which case computational efficiency incurs a significant penalty on the detection performance. Interestingly, this phenomenon is in line with the observation reported in Ma and Wu (2015) for the noisy submatrix detection problem, where the statistical limit can be attained if and only if the submatrix size exceeds the $(2/3)^{\text{th}}$ power of the matrix size.

4. Extensions and Open Problems

In this section, we discuss the extension of our results to: (1) the planted dense subgraph recovery and DKS problem; (2) the PDS problem where the planted dense subgraph has a deterministic size. (3) the bipartite PDS problem.

4.1. Recovering Planted Dense Subgraphs and DKS Problem

Closely related to the PDS detection problem is the recovery problem, where given a graph generated from $\mathcal{G}(N, K, p, q)$, the task is to recover the planted dense subgraph. As a consequence of our computational lower bound for detection, we discuss implications on the tractability of the recovery problem as well as the closely related DKS problem as illustrated in Fig. 2.

Consider the asymptotic regime of (1), where it has been shown Chen and Xu (2014); Ames (2013) that recovery is possible if and only if $\beta > \alpha$ and $\alpha < 1$. Note that in this case the recovery problem is harder than finding the DKS, because if the planted dense subgraph is recovered with high probability, we can obtain a $(1+\epsilon)$-approximation of the densest K-subgraph for any $\epsilon > 0$ in polynomial time.[2] Results in Chen and Xu (2014); Ames (2013) imply that the planted dense subgraph can be recovered in polynomial time in the simple (green) regime of Fig. 2 where $\beta > \frac{1}{2} + \frac{\alpha}{2}$. Consequently $(1+\epsilon)$-approximation of the DKS can be found efficiently in this regime.

Conversely, given a polynomial time η-factor approximation algorithm to the DKS problem with the output $\widehat{S}$, we can distinguish $H_0 : G \sim \mathcal{G}(N, q)$ versus $H_1 : G \sim \mathcal{G}(N, K, p = cq, q)$ if $\beta > \alpha$ and $c > \eta$ in polynomial time as follows: Fix any positive $\epsilon > 0$ such that $(1-\epsilon)c > (1+\epsilon)\eta$. Declare H_1 if the density of $\widehat{S}$ is larger than $(1+\epsilon)q$ and H_0 otherwise. Assuming $\beta > \alpha$, one can show that the density of $\widehat{S}$ is at most $(1+\epsilon)q$ under H_0 and at least $(1-\epsilon)p/\eta$ under H_1. Hence, our computational lower bounds for the PC problem imply that the densest K-subgraph as well as the planted dense subgraph is hard to approximate to any constant factor if $\alpha < \beta < \beta^\sharp(\alpha)$ (the red regime in Fig. 1). Whether DKS is hard to approximate with any constant factor in the blue regime of $\beta^\sharp(\alpha) \vee \alpha \le \beta \le \frac{1}{2} + \frac{\alpha}{2}$ is left as an interesting open problem.

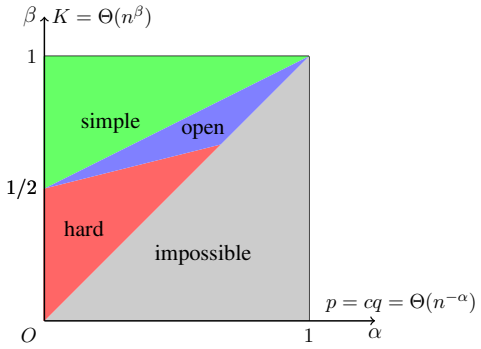

Figure 2: The simple (green), hard (red), impossible (gray) regimes for **recovering** planted dense subgraphs, and the hardness in the blue regime remains open.

4.2. PDS Problem with a Deterministic Size

In the PDS problem with a deterministic size K, the null distribution corresponds to the Erdős-Rényi graph $G(N, q)$; under the alternative, we choose K vertices uniformly at random to plant a dense subgraph with edge probability p. Although the subgraph size under our PDS model is binomially distributed, which, in the asymptotic regime (1), is sharply concentrated near its mean

2. If the planted dense subgraph size is smaller than K, output any K-subgraph containing it; otherwise output any of its K-subgraph.

K, it is not entirely clear whether these two models are equivalent. Although our reduction scheme in Section 3 extends to the fixed-size model with V_1^n being the random ℓ-partition of $[N]$ with $|V_t| = \ell$ for all $t \in [n]$, so far we have not been able to prove the alternative distributions are approximately matched: The main technical hurdle lies in controlling the total variation between the distribution of $\{E(V_t, V_t), t \in [n]\}$ after averaging over the random ℓ-partition $\{V_t\}$ and the desired distribution.

Nonetheless, our result on the hardness of solving the PDS problem extends to the case of deterministic dense subgraph size if the tests are required to be monotone. (A test ϕ is monotone if $\phi(G) = 1$ implies $\phi(G') = 1$ whenever G' is obtained by adding edges to G.) It is intuitive to assume that any reasonable test should be more likely to declare the existence of the planted dense subgraph if the graph contains more edges, such as the linear and scan test defined in (3). Moreover, by the monotonicity of the likelihood ratio, the statistically optimal test is also monotone. If we restrict our scope to monotone tests, then our computational lower bound implies that for the PDS problem with a deterministic size, there is no efficiently computable monotone test in the hard regime of $\alpha < \beta < \beta^\sharp$ in Fig. 1. In fact, for a given monotone polynomial-time solver ϕ for the PDS problem with size K, the $\mathsf{PDS}(N, 2K, p, q)$ can be solved by ϕ in polynomial time because with high probability the planted dense subgraph is of size at least K. We conjecture that the computational limit of PDS of fixed size is identical to that of the random size, which can indeed by established in the bipartite case as discussed in the next subsection. Also, we can show that the PDS *recovery* problem with a deterministic planted dense subgraph size K is computationally intractable if $\alpha < \beta < \beta^\sharp(\alpha)$ (the red regime in Fig. 1). See Appendix B for a formal statement and the proof.

4.3. Bipartite PDS Problem

Let $\mathcal{G}_{\mathrm{b}}(N, q)$ denote the bipartite Erdős-Rényi random graph model with N top vertices and N bottom vertices. Let $\mathcal{G}_{\mathrm{b}}(N, K, p, q)$ denote the bipartite variant of the planted densest subgraph model in Definition 1 with a planted dense subgraph of K top vertices and K bottom vertices on average. The bipartite PDS problem with parameters (N, K, p, q), denoted by $\mathsf{BPDS}(N, K, p, q)$, refers to the problem of testing $H_0 : G \sim \mathcal{G}_{\mathrm{b}}(N, q)$ versus $H_1 : G \sim \mathcal{G}_{\mathrm{b}}(N, K, p, q)$.

Consider the asymptotic regime of (1). Following the arguments in Section 2, one can show that the statistical limit is given by β^* defined in (4). To derive computational lower bounds, we use the reduction from the bipartite PC problem with parameters (n, k, γ), denoted by $\mathsf{BPC}(n, k, \gamma)$, which tests $H_0 : G \sim \mathcal{G}_{\mathrm{b}}(n, \gamma)$ versus $H_1 : G \sim \mathcal{G}_{\mathrm{b}}(n, k, \gamma)$, where $\mathcal{G}_{\mathrm{b}}(n, k, \gamma)$ is the bipartite variant of the planted clique model with a planted bi-clique of size $k \times k$. The BPC Hypothesis refers to the assumption that for some constant $0 < \gamma \le 1/2$, no sequence of randomized polynomial-time tests for BPC succeeds if $\limsup_{n\to\infty} \frac{\log k_n}{\log n} < 1/2$. The reduction scheme from $\mathsf{BPC}(n, k, \gamma)$ to $\mathsf{BPDS}(N, K, 2q, q)$ is analogue to the scheme used in non-bipartite case. The proof of computational lower bounds in bipartite graph is much simpler. In particular, under the null hypothesis, $G \sim \mathcal{G}_{\mathrm{b}}(n, \gamma)$ and one can verify that $\widetilde{G} \sim \mathcal{G}_{\mathrm{b}}(N, q)$. Under the alternative hypothesis, $G \sim \mathcal{G}_{\mathrm{b}}(n, k, \gamma)$. Lemma 8 directly implies that the total variation distance between the distribution of $\widetilde{G}$ and $\mathcal{G}_{\mathrm{b}}(N, K, 2q, q)$ is on the order of $k^2(q\ell^2)^{(m_0+1)}$. Then, following the arguments in Proposition 6 and Theorem 7, we conclude that if the BPC Hypothesis holds for any positive γ, then no efficiently computable test can solve $\mathsf{BPDS}(N, K, 2q, q)$ in the regime $\alpha < \beta < \beta^\sharp(\alpha)$ given by (8). The same conclusion also carries over to the bipartite PDS problem with a deterministic size K and the statistical and computational limits shown in Fig. 1 apply verbatim.

References

E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *Arxiv preprint arXiv:1405.3267*, 2014.

N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4), 1998.

N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 496–505. ACM, 2007.

N. Alon, S. Arora, R. Manokaran, D. Moshkovitz, and O. Weinstein. Inapproximabilty of densest κ-subgraph from average case hardness. *Manuscript, available at* https://www.nada.kth.se/~rajsekar/papers/dks.pdf, 2011.

B. PW Ames. Robust convex relaxation for the planted clique and densest k-subgraph problems. *arXiv:1305.4891*, 2013.

B. PW Ames and S. A Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical programming*, 129(1):69–89, 2011.

B. Applebaum, B. Barak, and A. Wigderson. Public-key cryptography from different assumptions. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 171–180, 2010. http://www.cs.princeton.edu/~boaz/Papers/ncpkcFull1.pdf.

E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 06 2014.

S. Arora, B. Barak, M. Brunnermeier, and R. Ge. Computational complexity and information asymmetry in financial products. In *Innovations in Computer Science (ICS 2010)*, pages 49–65, 2010. http://www.cs.princeton.edu/~rongge/derivativelatest.pdf.

S. Balakrishnan, M. Kolar, A. Rinaldo, A. Singh, and L. Wasserman. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, 2011.

Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res.*, 30:1046–1066 (electronic), 2013.

A. Bhaskara, M. Charikar, E. Chlamtac, U. Feige, and A. Vijayaraghavan. Detecting high log-densities: An $o(n^{1/4})$ approximation for densest k-subgraph. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 201–210, 2010.

C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 11 2013.

V. Chandrasekaran and M. I Jordan. Computational and statistical tradeoffs via convex relaxation. *PNAS*, 110(13):E1181–E1190, 2013.

Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv:1402.1267*, 2014.

Y. Dekel, O. Gurel-Gurevich, and Y. Peres. Finding hidden cliques in linear time with high probability. *arxiv:1010.2997*, 2010.

Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *arxiv:1304.7047*, 2012.

D. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13(2):99–124, 1998.

U. Feige and R. Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures & Algorithms*, 16(2):195–208, 2000.

U. Feige and D. Ron. Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA10), Discrete Math. Theor. Comput. Sci. Proc., AM*, pages 189–203, 2010.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pages 655–664, 2013.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *preprint, arxiv:1412.6156*, Nov 2014.

E. Hazan and R. Krauthgamer. How hard is it to approximate the best Nash equilibrium? *SIAM Journal on Computing*, 40(1):79–91, 2011.

M. Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4): 347–359, 1992.

A. Juels and M. Peinado. Hiding cliques for cryptographic security. *Designs, Codes & Crypto.*, 2000.

M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In *NIPS*, 2011.

L. Kučera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57 (2):193–212, 1995.

Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. to appear in *The Annals of Statistics, arXiv:1309.5914*, 2015.

L. Massoulié. Community detection thresholds and the weak Ramanujan property. *arxiv:1109.3318*, 2013.

F. McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529 – 537, 2001.

M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *available at: http://arxiv.org/abs/1202.1499*, 2012.

E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arxiv:1311.4115*, 2013.

E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *Arxiv preprint arXiv:1407.1591*, 2014.

R. Vershynin. A simple decoupling inequality in probability theory. Manuscript, available at http://www-personal.umich.edu/~romanv/papers/decoupling-simple.pdf, 2011.

N. Verzelen and E. Arias-Castro. Community detection in sparse random networks. *arXiv:1308.2955*, 2013.

J. Xu, R. Wu, K. Zhu, B. Hajek, R. Srikant, and L. Ying. Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs. *SIGMETRICS Perform. Eval. Rev.*, 42(1):29–41, June 2014.

Appendix A. Proofs

A.1. Proof of Proposition 3

Proof Let $\mathbb{P}_{A||S|}$ denote the distribution of A conditional on $|S|$ under the alternative hypothesis. Since $|S| \sim \mathrm{Binom}(N, K/N)$, by the Chernoff bound, $\mathbb{P}[|S| > 2K] \le \exp(-K/8)$. Therefore,

$$\begin{aligned} d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_1) &= d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{E}_{|S|}[\mathbb{P}_{A||S|}]) \\ &\le \mathbb{E}_{|S|}\left[d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_{A||S|})\right] \\ &\le \exp(-K/8) + \sum_{K' \le 2K} d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_{A||S|=K'})\mathbb{P}[|S| = K'], \end{aligned} \tag{9}$$

where the first inequality follows from the convexity of $(P, Q) \mapsto d_{\mathrm{TV}}(P, Q)$, Next we condition on $|S| = K'$ for a fixed $K' \le 2K$. Then S is uniformly distributed over all subsets of size K'. Let $\widetilde{S}$ be an independent copy of S. Then $|S \cap \widetilde{S}| \sim \mathrm{Hypergeometric}(N, K', K')$. By the definition of

the χ^2-divergence and Fubini's theorem,

$$\begin{aligned}
\chi^2(\mathbb{P}_{A||S|=K'}\|\mathbb{P}_0) &= \int \frac{\mathbb{E}_S[P_{A|S}]\mathbb{E}_{\widetilde{S}}[P_{A|\widetilde{S}}]}{\mathbb{P}_0} - 1 \\
&= \mathbb{E}_{S\perp\!\!\!\perp\widetilde{S}}\left[\int \frac{P_{A|S}P_{A|\widetilde{S}}}{\mathbb{P}_0}\right] - 1 \\
&= \mathbb{E}_{S\perp\!\!\!\perp\widetilde{S}}\left[\left(1+\frac{(p-q)^2}{q(1-q)}\right)^{\binom{|S\cap\widetilde{S}|}{2}}\right] - 1 \\
&\le \mathbb{E}_{S\perp\!\!\!\perp\widetilde{S}}\left[\exp\left(\frac{(c-1)^2 q}{1-q}\binom{|S\cap\widetilde{S}|}{2}\right)\right] - 1 \\
&\overset{(a)}{\le} \mathbb{E}\left[\exp\left((c-1)cq|S\cap\widetilde{S}|^2\right)\right] - 1 \\
&\overset{(b)}{\le} \tau(Cc^2) - 1,
\end{aligned}$$

where (a) is due to the fact that $q = \frac{p}{c} \le \frac{1}{c}$; (b) follows from Lemma 14 in Appendix C with an appropriate choice of function $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\tau(0+) = 1$. Therefore, we get that

$$2d_{\mathrm{TV}}^2(\mathbb{P}_0, \mathbb{P}_{A||S|=K'}) \le \log(\chi^2(\mathbb{P}_{A||S|=K'}\|\mathbb{P}_0) + 1) \le \log(\tau(Cc^2)), \tag{10}$$

Combining (9) and (10) yields (2) with $h \triangleq \log \circ \tau$. ■

A.2. Proof of Proposition 4

Proof Let $C > 0$ denote a constant whose value only depends on c and may change line by line. Under $\mathbb{P}_0$, $T_{\mathrm{lin}} \sim \mathrm{Binom}\left(\binom{N}{2}, q\right)$. By the Bernstein inequality,

$$\mathbb{P}_0[T_{\mathrm{lin}} > \tau_1] \le \exp\left(-\frac{\binom{K}{2}^2(p-q)^2/4}{2\binom{N}{2}q + \binom{K}{2}(p-q)/3}\right) \le \exp\left(-C\frac{K^4 q}{N^2}\right).$$

Under $\mathbb{P}_1$, Since $|S| \sim \mathrm{Binom}(N, K/N)$, by the Chernoff bound, $\mathbb{P}_1[|S| < 0.9K] \le \exp(-K/200)$. Conditional on $|S| = K'$ for some $K' \ge 0.9K$, then T_{lin} is distributed as an independent sum of $\mathrm{Binom}\left(\binom{K'}{2}, p\right)$ and $\mathrm{Binom}\left(\binom{N}{2} - \binom{K'}{2}, q\right)$. By the multiplicative Chernoff bound (see, e.g., (Mitzenmacher and Upfal, 2005, Theorem 4.5)),

$$\begin{aligned}
\mathbb{P}_1[T_{\mathrm{lin}} \le \tau_1] &\le \mathbb{P}_1[|S| < 0.9K] + \exp\left(-\frac{\left(2\binom{K'}{2} - \binom{K}{2}\right)^2(p-q)^2}{8\left(\binom{N}{2}q + \binom{K'}{2}(p-q)\right)}\right) \\
&\le \exp\left(-\frac{K}{200}\right) + \exp\left(-C\frac{K^4 q}{N^2}\right).
\end{aligned}$$

For the scan test statistic, under the null hypothesis, for any fixed subset S of size K, $\sum_{i,j\in S} A_{ij} \sim \mathrm{Binom}\left(\binom{K}{2}, q\right)$. By the union bound and the Bernstein inequality,

$$\mathbb{P}_0[T_{\mathrm{scan}} > \tau_2] \le \binom{N}{K}\mathbb{P}_0[\sum_{1\le i<j\le K} A_{ij} > \tau_2] \le \left(\frac{Ne}{K}\right)^K \exp\left(-\frac{\binom{K}{2}^2(p-q)^2/4}{2\binom{K}{2}q + \binom{K}{2}(p-q)/3}\right)$$
$$\le \exp\left(K\log\frac{Ne}{K} - CK^2q\right).$$

Under the alternative hypothesis, conditional on $|S| = K'$ for some $K' \ge 0.9K$, $\sum_{i,j\in S} A_{ij} \sim \mathrm{Binom}\left(\binom{K'}{2}, p\right)$ and thus T_{scan} is stochastically dominated by $\mathrm{Binom}\left(\binom{K'\wedge K}{2}, p\right)$. By the multiplicative Chernoff bound,

$$\mathbb{P}_1[T_{\mathrm{scan}} \le \tau_2] \le \mathbb{P}_1[|S| < 0.9K] + \exp\left(-\frac{\left(2\binom{K'\wedge K}{2} - \binom{K}{2}\right)^2(p-q)^2}{8\binom{K'\wedge K}{2}p}\right)$$
$$\le \exp\left(-\frac{K}{200}\right) + \exp\left(-CK^2q\right).$$

■

A.3. Proof of Proposition 5

We first introduce several key auxiliary results used in the proof. The following lemma ensures that $P'_{\ell_s\ell_t}$ and $Q'_{\ell_s\ell_t}$ are well-defined under suitable conditions and that $P'_{\ell_s\ell_t}$ and P_{ℓ_s,ℓ_t} are close in total variation.

Lemma 8 *Suppose that $p = 2q$ and $16q\ell^2 \le 1$. Fix $\{\ell_t\}$ such that $\ell_t \le 2\ell$ for all $t \in [k]$. Then for all $1 \le s < t \le k$, $P'_{\ell_s\ell_t}$ and $Q'_{\ell_s\ell_t}$ are probability measures and*

$$d_{\mathrm{TV}}(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}) \le 4(8q\ell^2)^{(m_0+1)}.$$

Proof Fix an (s,t) such that $1 \le s < t \le k$. We first show that $P'_{\ell_s\ell_t}$ and $Q'_{\ell_s\ell_t}$ are well-defined. By definition, $\sum_{m=0}^{\ell_s\ell_t} P'_{\ell_s\ell_t}(m) = \sum_{m=0}^{\ell_s\ell_t} Q'_{\ell_s\ell_t}(m) = 1$ and it suffices to show positivity, i.e.,

$$P_{\ell_s\ell_t}(0) + a_{\ell_s\ell_t} \ge 0, \tag{11}$$
$$Q_{\ell_s\ell_t}(m) \ge \gamma P'_{\ell_s\ell_t}(m), \quad \forall 0 \le m \le m_0. \tag{12}$$

Recall that $P_{\ell_s\ell_t} \sim \mathrm{Binom}(\ell_s\ell_t, p)$ and $Q_{\ell_s\ell_t} \sim \mathrm{Binom}(\ell_s\ell_t, q)$. Therefore,

$$Q_{\ell_s\ell_t}(m) = \binom{\ell_s\ell_t}{m}q^m(1-q)^{\ell_s\ell_t-m}, \quad P_{\ell_s\ell_t}(m) = \binom{\ell_s\ell_t}{m}p^m(1-p)^{\ell_s\ell_t-m}, \ \forall 0 \le m \le \ell_s\ell_t,$$

It follows that

$$\frac{1}{\gamma}Q_{\ell_s\ell_t}(m) - P_{\ell_s\ell_t}(m) = \frac{1}{\gamma}\binom{\ell_s\ell_t}{m}q^m(1-2q)^{\ell_s\ell_t-m}\left[\left(\frac{1-q}{1-2q}\right)^{\ell_s\ell_t-m} - 2^m\gamma\right].$$

Recall that $m_0 = \lfloor \log_2(1/\gamma) \rfloor$ and thus $Q_{\ell_s\ell_t}(m) \geq \gamma P_{\ell_s\ell_t}(m)$ for all $m \leq m_0$. Furthermore,

$$Q_{\ell_s\ell_t}(0) = (1-q)^{\ell_s\ell_t} \geq (1 - q\ell_s\ell_t) \geq 1 - 4q\ell^2 \geq \frac{3}{4} \geq \gamma \geq \gamma P'_{\ell_s\ell_t}(0),$$

and thus (12) holds. Recall that

$$a_{\ell_s\ell_t} = \sum_{m_0 < m \leq \ell_s\ell_t} \left(P_{\ell_s\ell_t}(m) - \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) \right)$$

Since $2^{m_0+1}\gamma > 1$ and $8q\ell^2 \leq 1/2$, it follows that

$$\frac{1}{\gamma} \sum_{m_0 < m \leq \ell_s\ell_t} Q_{\ell_s\ell_t}(m) \leq \frac{1}{\gamma} \sum_{m_0 < m \leq \ell_s\ell_t} \binom{\ell_s\ell_t}{m} q^m \leq \sum_{m > m_0} (2\ell_s\ell_t q)^m \leq 2(8q\ell^2)^{(m_0+1)}, \tag{13}$$

and therefore $a_{\ell_s\ell_t} \geq -1/2$. Furthermore,

$$P_{\ell_s\ell_t}(0) = (1-p)^{\ell_s\ell_t} \geq 1 - p\ell_s\ell_t \geq 1 - 8q\ell^2 \geq 1/2,$$

and thus (11) holds.

Next we bound $d_{\mathrm{TV}}\left(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}\right)$. Notice that

$$\sum_{m_0 < m \leq \ell_s\ell_t} P_{\ell_s\ell_t}(m) \leq \sum_{m_0 < m \leq \ell_s\ell_t} \binom{\ell_s\ell_t}{m} p^m \leq \sum_{m > m_0} (\ell_s\ell_t p)^m \leq 2(8q\ell^2)^{(m_0+1)}. \tag{14}$$

Therefore, by the definition of the total variation distance and $a_{\ell_s\ell_t}$,

$$\begin{aligned} d_{\mathrm{TV}}(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}) &= \frac{1}{2}|a_{\ell_s\ell_t}| + \frac{1}{2} \sum_{m_0 < m \leq \ell_s\ell_t} \left| P_{\ell_s\ell_t}(m) - \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) \right| \\ &\leq \sum_{m_0 < m \leq \ell_s\ell_t} \left(P_{\ell_s\ell_t}(m) + \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) \right) \leq 4(8q\ell^2)^{(m_0+1)}, \end{aligned}$$

where the last inequality follows from (13) and (14). ■

The following lemma is useful for upper bounding the total variation distance between a truncated mixture of product distribution P_Y and a product distribution Q_Y.

Lemma 9 *Let $P_{Y|X}$ be a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$ and denote the marginal of Y by $P_Y = \mathbb{E}_{X \sim P_X}[P_{Y|X}]$. Let Q_Y be such that $P_{Y|X=x} \ll Q_Y$ for all x. Let E be a measurable subset of $\mathcal{X}$. Define $g : \mathcal{X}^2 \to \bar{\mathbb{R}}_+$ by*

$$g(x, \widetilde{x}) \triangleq \int \frac{\mathrm{d}P_{Y|X=x} \mathrm{d}P_{Y|X=\widetilde{x}}}{\mathrm{d}Q}.$$

Then

$$d_{\mathrm{TV}}(P_Y, Q_Y) \leq \frac{1}{2} P_X(E^{\mathrm{c}}) + \frac{1}{2} \sqrt{\mathbb{E}\left[g(X, \widetilde{X}) \mathbf{1}_E(X) \mathbf{1}_E(\widetilde{X}) \right] - 1 + 2P_X(E^{\mathrm{c}})}, \tag{15}$$

where $\widetilde{X}$ is an independent copy of $X \sim P_X$.

Proof By definition of the total variation distance,

$$d_{\mathrm{TV}}(P_Y, Q_Y) = \frac{1}{2}\|P_Y - Q_Y\|_1 \le \frac{1}{2}\|\mathbb{E}[P_{Y|X}] - \mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}]\|_1 + \frac{1}{2}\|\mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}] - Q_Y\|_1,$$

where the first term is $\|\mathbb{E}[P_{Y|X}] - \mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}]\|_1 = \|\mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\notin E\}}]\|_1 = \mathbb{P}\{X \notin E\}$. The second term is controlled by

$$\begin{aligned}
\|\mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}] - Q_Y\|_1^2 &= \left(\mathbb{E}_{Q_Y}\left[\left|\frac{\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]}{Q_Y} - 1\right|\right]\right)^2 \\
&\le \mathbb{E}_{Q_Y}\left[\left(\frac{\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]}{Q_Y} - 1\right)^2\right] \qquad (16)\\
&= \mathbb{E}_{Q_Y}\left[\left(\frac{\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]}{Q_Y}\right)^2\right] + 1 - 2\,\mathbb{E}[\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]] \qquad (17)\\
&= \mathbb{E}\left[g(X,\tilde{X})\mathbf{1}_E(X)\mathbf{1}_E(\widetilde{X})\right] + 1 - 2\,\mathbb{P}\{X\in E\}, \qquad (18)
\end{aligned}$$

where (16) is Cauchy-Schwartz inequality, (18) follows from Fubini theorem. This proves the desired (15). ■

Note that $\{V_t : t \in [n]\}$ can be equivalently generated as follows: Throw balls indexed by $[N]$ into bins indexed by $[n]$ independently and uniformly at random; let V_t denote the set of balls in the t^{th} bin. Furthermore, Fix a subset $C \subset [n]$ and let $S = \cup_{t\in C} V_t$. Conditioned on S, $\{V_t : t \in C\}$ can be generated by throwing balls indexed by S into bins indexed by C independently and uniformly at random. We need the following negative association property (Dubhashi and Ranjan, 1998, Definition 1).

Lemma 10 *Fix a subset $C \subset [n]$ and let $S = \cup_{t\in C} V_t$. Let $\{\widetilde{V}_t : t \in C\}$ be an independent copy of $\{V_t : t \in C\}$ conditioned on S. Then conditioned on S, the full vector $\{|V_s \cap \widetilde{V}_t| : s, t \in C\}$ is negatively associated, i.e., for every two disjoint index sets $I, J \subset C \times C$,*

$$\mathbb{E}[f(V_s \cap \widetilde{V}_t, (s,t) \in I)g(V_s \cap \widetilde{V}_t, (s,t) \in J)] \le \mathbb{E}[f(V_s \cap \widetilde{V}_t, (s,t) \in I)]\mathbb{E}[g(V_s \cap \widetilde{V}_t, (s,t) \in J)],$$

for all functions $f : \mathbb{R}^{|I|} \to \mathbb{R}$ and $g : \mathbb{R}^{|J|} \to \mathbb{R}$ that are either both non-decreasing or both non-increasing in every argument.

Proof Define the indicator random variables $Z_{m,s,t}$ for $m \in S, s, t \in C$ as

$$Z_{m,s,t} = \begin{cases} 1 & \text{if the } m^{\text{th}} \text{ ball is contained in } V_s \text{ and } \widetilde{V}_t, \\ 0 & \text{otherwise .} \end{cases}$$

By (Dubhashi and Ranjan, 1998, Proposition 12), the full vector $\{Z_{m,s,t} : m \in S, s, t \in C\}$ is negatively associated. By definition, we have

$$|V_s \cap \widetilde{V}_t| = \sum_{m\in S} Z_{m,s,t},$$

which is a non-decreasing function of $\{Z_{m,s,t} : m \in S\}$. Moreover, for distinct pairs $(s,t) \neq (s',t')$, the sets $\{(m,s,t) : m \in S\}$ and $\{(m,s',t') : m \in S\}$ are disjoint. Applying (Dubhashi and Ranjan, 1998, Proposition 8) yields the desired statement. ■

The negative association property of $\{|V_s \cap \widetilde{V}_t| : s,t \in C\}$ allows us to bound the expectation of any non-decreasing function of $\{|V_s \cap \widetilde{V}_t| : s,t \in C\}$ conditional on C and S as if they were independent (Dubhashi and Ranjan, 1998, Lemma 2), i.e., for any collection of non-decreasing functions $\{f_{s,t} : s,t \in [n]\}$,

$$\mathbb{E}\left[\prod_{s,t\in C} f_{s,t}(|V_s \cap \widetilde{V}_t|) \,\middle|\, C, S\right] \leq \prod_{s,t\in C} \mathbb{E}\left[f_{s,t}(|V_s \cap \widetilde{V}_t|) \,\middle|\, C, S\right]. \tag{19}$$

Lemma 11 *Suppose that* $X \sim \mathrm{Binom}(1.5K, \frac{1}{k^2})$ *and* $Y \sim \mathrm{Binom}(3\ell, \frac{e}{k})$ *with* $K = k\ell$ *and* $k \geq 6e\ell$. *Then for all* $1 \leq m \leq 2\ell - 1$,

$$\mathbb{P}[X = m] \leq \mathbb{P}[Y = m],$$

and $\mathbb{P}[X \geq 2\ell] \leq \mathbb{P}[Y = 2\ell]$.

Proof In view of the fact that $(\frac{n}{m})^m \leq \binom{n}{m} \leq (\frac{en}{m})^m$, we have for $1 \leq m \leq 2\ell$,

$$\mathbb{P}[X = m] = \binom{1.5K}{m}\left(\frac{1}{k^2}\right)^m\left(1 - \frac{1}{k^2}\right)^{1.5K-m} \leq \left(\frac{1.5eK}{mk^2}\right)^m.$$

Therefore,

$$\mathbb{P}[X \geq 2\ell] \leq \sum_{m=2\ell}^{\infty}\left(\frac{1.5e\ell}{km}\right)^m \leq \sum_{m=2\ell}^{\infty}\left(\frac{3e}{4k}\right)^m \leq \frac{(0.75e/k)^{2\ell}}{1 - 0.75e/k}.$$

On the other hand, for $1 \leq m \leq 2\ell - 1$

$$\begin{aligned}\mathbb{P}[Y = m] &= \binom{3\ell}{m}\left(\frac{e}{k}\right)^m\left(1 - \frac{e}{k}\right)^{3\ell-m} \\ &\geq \left(\frac{3e\ell}{mk}\right)^m\left(1 - \frac{3e\ell}{k}\right) \\ &\geq 2^{m-1}\left(\frac{1.5e\ell}{mk}\right)^m \geq \mathbb{P}[X = m].\end{aligned}$$

Moreover, $\mathbb{P}[Y = 2\ell] \geq \mathbb{P}[X \geq 2\ell]$. ■

Lemma 12 *Let* $T \sim \mathrm{Binom}(\ell, \tau)$ *and* $\lambda > 0$. *Assume that* $\lambda\ell \leq \frac{1}{16}$. *Then*

$$\mathbb{E}[\exp(\lambda T(T-1))] \leq \exp\left(16\lambda\ell^2\tau^2\right). \tag{20}$$

Proof Let $(s_1, \ldots, s_\ell, t_1, \ldots, t_\ell) \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(\tau)$, $S = \sum_{i=1}^{\ell} s_i$ and $T = \sum_{i=1}^{\ell} t_i$. Next we use a decoupling argument to replace $T^2 - T$ by ST:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda T(T-1)\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i \neq j} t_i t_j\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(4\lambda \sum_{i \neq j} s_i t_j\right)\right], \qquad (21) \\
&\leq \mathbb{E}\left[\exp\left(4\lambda ST\right)\right],
\end{aligned}$$

where (21) is a standard decoupling inequality (see, e.g., (Vershynin, 2011, Theorem 1)). Since $\lambda T \leq \lambda \ell \leq \frac{1}{16}$ and $\exp(x) - 1 \leq \exp(a)x$ for all $x \in [0, a]$, the desired (20) follows from

$$\begin{aligned}
\mathbb{E}\left[\exp\left(4\lambda ST\right)\right] &= \mathbb{E}\left[\left(1 + \tau(\exp(4\lambda T) - 1)\right)^{\ell}\right] \\
&\leq \mathbb{E}\left[\left(1 + 8\tau\lambda T\right)^{\ell}\right] \\
&\leq \mathbb{E}\left[\exp\left(8\tau\lambda\ell T\right)\right] \\
&= \left(1 + \tau\left(\exp\left(8\tau\lambda\ell\right) - 1\right)\right)^{\ell} \\
&\leq \exp\left(16\tau^2\lambda\ell^2\right).
\end{aligned}$$

■

Proof [Proof of Proposition 5] Let $[i, j]$ denote the unordered pair of i and j. For any set $I \subset [N]$, let $\mathcal{E}(I)$ denote the set of unordered pairs of distinct elements in I, i.e., $\mathcal{E}(I) = \{[i, j] : i, j \in S, i \neq j\}$, and let $\mathcal{E}(I)^c = \mathcal{E}([N]) \setminus \mathcal{E}(I)$. For $s, t \in [n]$ with $s \neq t$, let $\widetilde{G}_{V_s V_t}$ denote the bipartite graph where the set of left (right) vertices is V_s (resp. V_t) and the set of edges is the set of edges in $\widetilde{G}$ from vertices in V_s to vertices in V_t. For $s \in [n]$, let $\widetilde{G}_{V_s V_s}$ denote the subgraph of $\widetilde{G}$ induced by V_s. Let $\widetilde{P}_{V_s V_t}$ denote the edge distribution of $\widetilde{G}_{V_s V_t}$ for $s, t \in [n]$.

First, we show that the null distributions are exactly matched by the reduction scheme. Lemma 8 implies that $P'_{\ell_s \ell_t}$ and $Q'_{\ell_s \ell_t}$ are well-defined probability measures, and by definition, $(1-\gamma)Q'_{\ell_s \ell_t} + \gamma P'_{\ell_s \ell_t} = Q_{\ell_s \ell_t} = \mathrm{Binom}(\ell_s \ell_t, q)$. Under the null hypothesis, $G \sim \mathcal{G}(n, \gamma)$ and therefore, according to our reduction scheme, $E(V_s, V_t) \sim \mathrm{Binom}(\ell_s \ell_t, q)$ for $s < t$ and $E(V_t, V_t) \sim \mathrm{Binom}(\binom{\ell_t}{2}, q)$. Since the vertices in V_s and V_t are connected uniformly at random such that the total number of edges is $E(V_s, V_t)$, it follows that $\widetilde{P}_{V_s V_t} = \prod_{(i,j) \in V_s \times V_t} \mathrm{Bern}(q)$ for $s < t$ and $\widetilde{P}_{V_s V_t} = \prod_{[i,j] \in \mathcal{E}(V_s)} \mathrm{Bern}(q)$ for $s = t$. Conditional on V_1^n, $\{E(V_s, V_t) : 1 \leq s \leq t \leq n\}$ are independent and so are $\{\widetilde{G}_{V_s V_t} : 1 \leq s \leq t \leq n\}$. Consequently, $P_{\widetilde{G}|H_0^C} = \mathbb{P}_0 = \prod_{[i,j] \in \mathcal{E}([N])} \mathrm{Bern}(q)$ and $\widetilde{G} \sim \mathcal{G}(N, q)$.

Next, we proceed to consider the alternative hypothesis, under which G is drawn from the planted clique model $\mathcal{G}(n, k, \gamma)$. Let $C \subset [n]$ denote the planted clique. Define $S = \cup_{t \in C} V_t$ and recall $K = k\ell$. Then $|S| \sim \mathrm{Binom}(N, K/N)$ and conditional on $|S|$, S is uniformly distributed over all possible subsets of size $|S|$ in $[N]$. By the symmetry of the vertices of G, the distribution of $\widetilde{A}$ conditional on C does not depend on C. Hence, without loss of generality, we shall assume that

$C = [k]$ henceforth. The distribution of $\widetilde{A}$ can be written as a mixture distribution indexed by the random set S as

$$\widetilde{A} \sim \widetilde{\mathbb{P}}_1 \triangleq \mathbb{E}_S \left[\widetilde{P}_{SS} \times \prod_{[i,j] \in \mathcal{E}(S)^c} \mathrm{Bern}(q) \right],$$

By the definition of $\mathbb{P}_1$,

$$\begin{aligned}
d_{\mathrm{TV}}(\widetilde{\mathbb{P}}_1, \mathbb{P}_1) &= d_{\mathrm{TV}} \left(\mathbb{E}_S \left[\widetilde{P}_{SS} \times \prod_{[i,j] \in \mathcal{E}(S)^c} \mathrm{Bern}(q) \right], \mathbb{E}_S \left[\prod_{[i,j] \in \mathcal{E}(S)} \mathrm{Bern}(p) \prod_{[i,j] \in \mathcal{E}(S)^c} \mathrm{Bern}(q) \right] \right) \\
&\le \mathbb{E}_S \left[d_{\mathrm{TV}} \left(\widetilde{P}_{SS} \times \prod_{[i,j] \in \mathcal{E}(S)^c} \mathrm{Bern}(q), \prod_{[i,j] \in \mathcal{E}(S)} \mathrm{Bern}(p) \prod_{[i,j] \in \mathcal{E}(S)^c} \mathrm{Bern}(q) \right) \right] \\
&= \mathbb{E}_S \left[d_{\mathrm{TV}} \left(\widetilde{P}_{SS}, \prod_{[i,j] \in \mathcal{E}(S)} \mathrm{Bern}(p) \right) \right] \\
&\le \mathbb{E}_S \left[d_{\mathrm{TV}} \left(\widetilde{P}_{SS}, \prod_{[i,j] \in \mathcal{E}(S)} \mathrm{Bern}(p) \right) \mathbf{1}_{\{|S| \le 1.5K\}} \right] + \exp(-K/12), \qquad (22)
\end{aligned}$$

where the first inequality follows from the convexity of $(P, Q) \mapsto d_{\mathrm{TV}}(P, Q)$, and the last inequality follows from applying the Chernoff bound to $|S|$. Fix an $S \subset [N]$ such that $|S| \le 1.5K$. Define $P_{V_t V_t} = \prod_{[i,j] \in \mathcal{E}(V_t)} \mathrm{Bern}(q)$ for $t \in [k]$ and $P_{V_s V_t} = \prod_{(i,j) \in V_s \times V_t} \mathrm{Bern}(p)$ for $1 \le s < t \le k$. By the triangle inequality,

$$\begin{aligned}
d_{\mathrm{TV}} \left(\widetilde{P}_{SS}, \prod_{[i,j] \in \mathcal{E}(S)} \mathrm{Bern}(p) \right) &\le d_{\mathrm{TV}} \left(\widetilde{P}_{SS}, \mathbb{E}_{V_1^k} \left[\prod_{1 \le s \le t \le k} P_{V_s V_t} \,\middle|\, S \right] \right) \qquad (23) \\
&\quad + d_{\mathrm{TV}} \left(\mathbb{E}_{V_1^k} \left[\prod_{1 \le s \le t \le k} P_{V_s V_t} \,\middle|\, S \right], \prod_{[i,j] \in \mathcal{E}(S)} \mathrm{Bern}(p) \right). \qquad (24)
\end{aligned}$$

To bound the term in (23), first note that conditional on S, $\{V_1^k\}$ can be generated as follows: Throw balls indexed by S into bins indexed by $[k]$ independently and uniformly at random; let V_t is the set of balls in the t^{th} bin. Define the event $E = \{V_1^k : |V_t| \le 2\ell, t \in [k]\}$. Since $|V_t| \sim \mathrm{Binom}(|S|, 1/k)$ is stochastically dominated by $\mathrm{Binom}(1.5K, 1/k)$ for each fixed $1 \le t \le k$, it

follows from the Chernoff bound and the union bound that $\mathbb{P}\{E^c\} \le k\exp(-\ell/18)$.

$$\begin{aligned}
&d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right]\right)\\
&\overset{(a)}{=} d_{\mathrm{TV}}\left(\mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t} \,\middle|\, S\right], \mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right]\right)\\
&\le \mathbb{E}_{V_1^k}\left[d_{\mathrm{TV}}\left(\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s\le t\le k} P_{V_sV_t}\right) \,\middle|\, S\right]\\
&\le \mathbb{E}_{V_1^k}\left[d_{\mathrm{TV}}\left(\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s\le t\le k} P_{V_sV_t}\right)\mathbf{1}_{\{V_1^k\in E\}} \,\middle|\, S\right] + k\exp(-\ell/18),
\end{aligned}$$

where (a) holds because conditional on V_1^k, $\left\{\widetilde{A}_{V_sV_t} : s,t\in[k]\right\}$ are independent. Recall that $\ell_t = |V_t|$. For any fixed $V_1^k \in E$, we have

$$\begin{aligned}
d_{\mathrm{TV}}\left(\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s\le t\le k} P_{V_sV_t}\right) &\overset{(a)}{=} d_{\mathrm{TV}}\left(\prod_{1\le s<t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s<t\le k} P_{V_sV_t}\right)\\
&\overset{(b)}{=} d_{\mathrm{TV}}\left(\prod_{1\le s<t\le k} P'_{\ell_s\ell_t}, \prod_{1\le s<t\le k} P_{\ell_s\ell_t}\right)\\
&\le d_{\mathrm{TV}}\left(\prod_{1\le s<t\le k} P'_{\ell_s\ell_t}, \prod_{1\le s<t\le k} P_{\ell_s\ell_t}\right)\\
&\le \sum_{1\le s<t\le k} d_{\mathrm{TV}}\left(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}\right) \overset{(c)}{\le} 2k^2(8q\ell^2)^{(m_0+1)},
\end{aligned}$$

where (a) follows since $\widetilde{P}_{V_tV_t} = P_{V_tV_t}$ for all $t\in[k]$; (b) is because the number of edges $E(V_s,V_t)$ is a sufficient statistic for testing $\widetilde{P}_{V_sV_t}$ versus $P_{V_sV_t}$ on the submatrix $A_{V_sV_t}$ of the adjacency matrix; (c) follows from Lemma 8. Therefore,

$$d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right]\right) \le 2k^2(8q\ell^2)^{(m_0+1)} + k\exp(-\ell/18). \tag{25}$$

To bound the term in (24), applying Lemma 9 yields

$$\begin{aligned}
&d_{\mathrm{TV}}\left(\mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right], \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right)\\
&\le \frac{1}{2}\mathbb{P}\left\{E^c\right\} + \frac{1}{2}\sqrt{\mathbb{E}_{V_1^k,\widetilde{V}_1^k}\left[g(V_1^k,\widetilde{V}_1^k)\mathbf{1}_{\{V_1^k\in E\}}\mathbf{1}_{\{\widetilde{V}_1^k\in E\}} \,\middle|\, S\right] - 1 + 2\mathbb{P}\left\{E^c\right\}},
\end{aligned} \tag{26}$$

where

$$g(V_1^k, \widetilde{V}_1^k) = \int \frac{\prod_{1\le s\le t\le k} P_{V_sV_t} \prod_{1\le s\le t\le k} P_{\widetilde{V}_s\widetilde{V}_t}}{\prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)}$$
$$= \prod_{s,t=1}^{k} \left(\frac{q^2}{p} + \frac{(1-q)^2}{1-p}\right)^{\binom{|V_s\cap\widetilde{V}_t|}{2}}$$
$$= \prod_{s,t=1}^{k} \left(\frac{1-\frac{3}{2}q}{1-2q}\right)^{\binom{|V_s\cap\widetilde{V}_t|}{2}}.$$

Let $X \sim \mathrm{Bin}(1.5K, \frac{1}{k^2})$ and $Y \sim \mathrm{Bin}(3\ell, e/k)$. It follows that

$$\mathbb{E}_{V_1^k;\widetilde{V}_1^k}\left[\prod_{s,t=1}^{k}\left(\frac{1-\frac{3}{2}q}{1-2q}\right)^{\binom{|V_s\cap\widetilde{V}_t|}{2}} \prod_{s,t=1}^{k} \mathbf{1}_{\{|V_s|\le 2\ell, |\widetilde{V}_t|\le 2\ell\}} \,\middle|\, S\right]$$
$$\overset{(a)}{\le} \mathbb{E}_{V_1^k;\widetilde{V}_1^k}\left[\prod_{s,t=1}^{k} e^{q\binom{|V_s\cap\widetilde{V}_t|\wedge 2\ell}{2}} \,\middle|\, S\right]$$
$$\overset{(b)}{\le} \prod_{s,t=1}^{k} \mathbb{E}\left[e^{q\binom{|V_s\cap\widetilde{V}_t|\wedge 2\ell}{2}} \,\middle|\, S\right]$$
$$\overset{(c)}{\le} \left(\mathbb{E}\left[e^{q\binom{X\wedge 2\ell}{2}}\right]\right)^{k^2} \overset{(d)}{\le} \mathbb{E}\left[e^{q\binom{Y}{2}}\right]^{k^2} \overset{(e)}{\le} \exp(72e^2q\ell^2), \tag{27}$$

where (a) follows from $1 + x \le e^x$ for all $x \ge 0$ and $q < 1/4$; (b) follows from the negative association property of $\{|V_s \cap \widetilde{V}_t| : s, t \in [k]\}$ proved in Lemma 10 and (19), in view of the monotonicity of $x \mapsto e^{q\binom{x\wedge 2\ell}{2}}$ on $\mathbb{R}_+$; (c) follows because $|V_s \cap \widetilde{V}_t|$ is stochastically dominated by $\mathrm{Binom}(1.5K, 1/k^2)$ for all $(s, t) \in [k]^2$; (d) follows from Lemma 11; (e) follows from Lemma 12 with $\lambda = q/2$ and $q\ell \le 1/8$. Therefore, by (26)

$$d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right) \le 0.5ke^{-\frac{\ell}{18}} + 0.5\sqrt{e^{72e^2q\ell^2} - 1 + 2ke^{-\frac{\ell}{18}}}$$
$$\le 0.5ke^{-\frac{\ell}{18}} + 0.5\sqrt{e^{72e^2q\ell^2} - 1} + \sqrt{0.5k}e^{-\frac{\ell}{36}}. \tag{28}$$

The proposition follows by combining (22), (23), (24), (25) and (28). ■

A.4. Proof of Proposition 6

Proof By assumption the test ϕ satisfies

$$\mathbb{P}_0\{\phi(G') = 1\} + \mathbb{P}_1\{\phi(G') = 0\} = \eta,$$

where G' is the graph in $\mathsf{PDS}(N, K, 2q, q)$ distributed according to either $\mathbb{P}_0$ or $\mathbb{P}_1$. Let G denote the graph in the $\mathsf{PC}(n, k, \gamma)$ and $\widetilde{G}$ denote the corresponding output of the randomized reduction scheme. Proposition 5 implies that $\widetilde{G} \sim \mathcal{G}(N, q)$ under H_0^{C}. Therefore $\mathbb{P}_{H_0^{\mathrm{C}}}\{\phi(\widetilde{G}) = 1\} = \mathbb{P}_0\{\phi(G') = 1\}$. Moreover,

$$|\mathbb{P}_{H_1^{\mathrm{C}}}\{\phi(\widetilde{G}) = 0\} - \mathbb{P}_1\{\phi(G') = 0\}| \le d_{\mathrm{TV}}(P_{\widetilde{G}|H_1^C}, \mathbb{P}_1) \le \xi.$$

It follows that

$$\mathbb{P}_{H_0^{\mathrm{C}}}\{\phi(\widetilde{G}) = 1\} + \mathbb{P}_{H_1^{\mathrm{C}}}\{\phi(\widetilde{G}) = 0\} \le \eta + \xi.$$

■

A.5. Proof of Theorem 7

Proof Fix $\alpha > 0$ and $0 < \beta < 1$ that satisfy (7). Let $\delta = 2/(m_0\alpha)$. Then it is straightforward to verify that $\frac{2+m_0\delta}{4+2\delta}\alpha \le \frac{1}{2} - \delta + \frac{1+2\delta}{4+2\delta}\alpha$. It follows from the assumption (7) that

$$\alpha < \beta < \min\left\{\frac{2+m_0\delta}{4+2\delta}\alpha,\ \frac{1}{2} - \delta + \frac{1+2\delta}{4+2\delta}\alpha\right\}. \tag{29}$$

Let $\ell \in \mathbb{N}$ and $q_\ell = \ell^{-(2+\delta)}$. Define

$$n_\ell = \lfloor \ell^{\frac{2+\delta}{\alpha}-1} \rfloor,\ k_\ell = \lfloor \ell^{\frac{(2+\delta)\beta}{\alpha}-1} \rfloor,\ N_\ell = n_\ell\ell,\ K_\ell = k_\ell\ell. \tag{30}$$

Then

$$\lim_{\ell\to\infty} \frac{\log \frac{1}{q_\ell}}{\log N_\ell} = \frac{(2+\delta)}{(2+\delta)/\alpha - 1 + 1} = \alpha,\ \lim_{\ell\to\infty} \frac{\log K_\ell}{\log N_\ell} = \frac{(2+\delta)\beta/\alpha - 1 + 1}{(2+\delta)/\alpha - 1 + 1} = \beta. \tag{31}$$

Suppose that for the sake of contradiction there exists a small $\epsilon > 0$ and a sequence of randomized polynomial-time tests $\{\phi_\ell\}$ for $\mathsf{PDS}(N_\ell, K_\ell, 2q_\ell, q_\ell)$, such that

$$\mathbb{P}_0\{\phi_{N_\ell,K_\ell}(G') = 1\} + \mathbb{P}_1\{\phi_{N_\ell,K_\ell}(G') = 0\} \le 1 - \epsilon$$

holds for arbitrarily large ℓ, where G' is the graph in the $\mathsf{PDS}(N_\ell, K_\ell, 2q_\ell, q_\ell)$. Since $\beta > \alpha$, we have $k_\ell \ge \ell^{1+\delta}$. Therefore, $16q_\ell\ell^2 \le 1$ and $k_\ell \ge 6e\ell$ for all sufficiently large ℓ. Applying Proposition 6, we conclude that $G \mapsto \phi(\widetilde{G})$ is a randomized polynomial-time test for $\mathsf{PC}(n_\ell, k_\ell, \gamma)$ whose Type-I+II error probability satisfies

$$\mathbb{P}_{H_0^{\mathrm{C}}}\{\phi_\ell(\widetilde{G}) = 1\} + \mathbb{P}_{H_1^{\mathrm{C}}}\{\phi_\ell(\widetilde{G}) = 0\} \le 1 - \epsilon + \xi, \tag{32}$$

where ξ is given by the right-hand side of (6). By the definition of q_ℓ, we have $q_\ell\ell^2 = \ell^{-\delta}$ and thus

$$k_\ell^2(q_\ell\ell^2)^{m_0+1} \le \ell^{2((2+\delta)\beta/\alpha-1)-(m_0+1)\delta} \le \ell^{-\delta},$$

where the last inequality follows from (29). Therefore $\xi \to 0$ as $\ell \to \infty$. Moreover, by the definition in (30),

$$\lim_{\ell\to\infty} \frac{\log k_\ell}{\log n_\ell} = \frac{(2+\delta)\beta/\alpha - 1}{(2+\delta)/\alpha - 1} \le \frac{1}{2} - \delta,$$

where the above inequality follows from (29). Therefore, (32) contradicts our assumption that Hypothesis 1 holds for γ. Finally, if Hypothesis 1 holds for any $\gamma > 0$, (8) follows from (7) by sending $\gamma \downarrow 0$. ■

Appendix B. Computational Lower Bounds for Approximately Recovering a Planted Dense Subgraph with Deterministic Size

Let $\widetilde{\mathcal{G}}(N, K, p, q)$ denote the planted dense subgraph model with N vertices and a deterministic dense subgraph size K: (1) A random set S of size K is uniformly chosen from $[N]$; (2) for any two vertices, they are connected with probability p if both of them are in S and with probability q otherwise, where $p > q$. Let $\mathsf{PDSR}\ (n, K, p, q, \epsilon)$ denote the planted dense subgraph recovery problem, where given a graph generated from $\widetilde{\mathcal{G}}(N, K, p, q)$ and an $\epsilon < 1$, the task is to output a set $\widehat{S}$ of size K such that $\widehat{S}$ is a $(1-\epsilon)$-approximation of S, i.e., $|\widehat{S} \cap S| \ge (1-\epsilon)K$. The following theorem implies that $\mathsf{PDSR}\ (N, K, p = cq, q, \epsilon)$ is at least as hard as $\mathsf{PDS}\ (N, K, p = cq, q)$ if $Kq = \Omega(\log N)$. Notice that in $\mathsf{PDSR}\ (N, K, p, q, \epsilon)$, the planted dense subgraph has a deterministic size K, while in $\mathsf{PDS}\ (N, K, p, q)$, the size of the planted dense subgraph is binomially distributed with mean K.

Theorem 13 *For any constant $\epsilon < 1$ and $c > 0$, suppose there is an algorithm $\mathcal{A}_N$ with running time T_N that solves the $\mathsf{PDSR}\ (N, K, cq, q, \epsilon)$ problem with probability $1 - \eta_N$. Then there exists a test ϕ_N with running time at most $N^2 + NT_N + NK^2$ that solves the $\mathsf{PDS}\ (N, 2K, cq, q)$ problem with Type-I+II error probabilities at most $\eta_N + e^{-CK} + 2Ne^{-CK^2 q + K\log N}$, where the constant $C > 0$ only depends on ϵ and c.*

Proof Given a graph G, we construct a sequence of graphs $G_1, \ldots, G_N$ sequentially as follows: Choose a permutation π on the N vertices uniformly at random. Let $G_0 = G$. For each $t \in [N]$, replace the vertex $\pi(t)$ in G_{t-1} with a new vertex that connects to all other vertices independently at random with probability q. We run the given algorithm $\mathcal{A}_N$ on $G_1, \ldots, G_N$ and let $S_1, \ldots, S_N$ denote the outputs which are sets of K vertices. Let $E(S_i, S_i)$ denote the total number of edges in S_i and $\tau = q + (1-\epsilon)^2(p-q)/2$. Define a test $\phi : G \to \{0, 1\}$ such that $\phi(G) = 1$ if and only if $\max_{i\in[N]} E(S_i, S_i) > \tau\binom{K}{2}$. The construction of each G_i takes N time units; the running time of $\mathcal{A}$ on G_i is at most T_N time units; the computation of $E(S_i, S_i)$ takes at most K^2 time units. Therefore, the total running time of ϕ is at most $N^2 + NT_N + NK^2$.

Next we upper bound the Type-I and II error probabilities of ϕ. Let $C = C(\epsilon, c)$ denote a positive constant whose value may depend on the context. If $G \sim \mathcal{G}(N, q)$, then all G_i are distributed

according to $\mathcal{G}(N,q)$. By the union bound and the Bernstein inequality,

$$\begin{aligned}
\mathbb{P}_0\{\phi(G)=1\} &\le \sum_{i=1}^N \mathbb{P}_0\left\{E(S_i,S_i)\ge \tau\binom{K}{2}\right\}\\
&\le \sum_{i=1}^N \sum_{S':S'\subset[N],|S'|=K} \mathbb{P}_0\left\{E(S',S')\ge \tau\binom{K}{2}\right\}\\
&\le N\binom{N}{K}\exp\left(-\frac{\binom{K}{2}^2(1-\epsilon)^4(p-q)^2/4}{2\binom{K}{2}q+\binom{K}{2}(1-\epsilon)^2(p-q)/3}\right)\\
&\le N\exp(-CK^2q+K\log N).
\end{aligned}$$

If $G\sim\mathcal{G}(N,2K,p,q)$, let S denote the set of vertices in the planted dense subgraph. Then $|S|\sim \mathrm{Binom}(N,\frac{2K}{N})$ and by the Chernoff bound, $\mathbb{P}_1[|S|<K]\le\exp(-CK)$. If $|S|=K'\ge K$, then there must exist some $I\in[N]$ such that G_I is distributed exactly as $\widetilde{\mathcal{G}}(N,K,p,q)$. Let S^* denote the set of vertices in the planted dense subgraph of G_I such that $|S^*|=K$. Then conditional on $I=i$ and the success of $\mathcal{A}_N$ on G_i, $|S_i\cap S^*|\ge(1-\epsilon)K$. Thus by the union bound and the Bernstein inequality, for $K'\ge K$,

$$\begin{aligned}
&\mathbb{P}_1\{\phi(G)=0||S|=K',I=i\}\\
&\le \eta_N + \sum_{S'\subset[N]:|S'|=K,|S'\cap S^*|\ge(1-\epsilon)K} \mathbb{P}_1\left\{E(S',S')\le\tau\binom{K}{2}\middle||S|=K',I=i\right\}\\
&\le \eta_N + \sum_{t\ge(1-\epsilon)K}^K \binom{K}{t}\binom{N-K}{K-t}\exp\left(-\frac{\binom{K}{2}^2(1-\epsilon)^4(p-q)^2/4}{2\binom{K}{2}p+\binom{K}{2}(1-\epsilon)^2(p-q)/3}\right)\\
&\le \eta_N + K\exp(-CK^2q+K\log N).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{P}_1\{\phi(G)=0\}\\
&\le \mathbb{P}_1\{|S|<K\}+\sum_{K'\ge K}\sum_{i=1}^N \mathbb{P}_1\{|S|=K',I=i\}\mathbb{P}_1\{\phi(G)=0||S|=K',I=i\}\\
&\le \exp(-CK)+\eta_N+K\exp(-CK^2q+K\log N).
\end{aligned}$$

■

Appendix C. A Lemma on Hypergeometric Distributions

Lemma 14 *There exists a function $\tau:\mathbb{R}_+\to\mathbb{R}_+$ satisfying $\tau(0+)=1$ such that the following holds: For any $p\in\mathbb{N}$ and $m\in[p]$, let $H\sim$ Hypergeometric(p,m,m) and $\lambda=b\left(\frac{1}{m}\log\frac{ep}{m}\wedge\frac{p^2}{m^4}\right)$ with $0<b<1/(16\mathrm{e})$. Then*

$$\mathbb{E}\left[\exp\left(\lambda H^2\right)\right]\le\tau(b). \tag{33}$$

Proof Notice that if $p \le 64$, then the lemma trivially holds. Hence, assume $p \ge 64$ in the rest of the proof. We consider three separate cases depending on the value of m. We first deal with the case of $m \ge \frac{p}{4}$. Then $\lambda = \frac{bp^2}{m^4} \le \frac{256b}{p^2}$. Since $H \le p$ with probability 1, we have $\mathbb{E}\left[\exp\left(\lambda H^2\right)\right] \le \exp(256b)$.

Next assume that $m \le \log \frac{ep}{m}$. Then $m \le \log p$ and $\lambda = \frac{b}{m} \log \frac{ep}{m}$. Let $(s_1, \ldots, s_m) \overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{m}{p-m})$. Then $S = \sum_{i=1}^m s_i \sim \text{Bin}(m, \frac{m}{p-m})$ which dominates H stochastically. It follows that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda H^2\right)\right] &\le \mathbb{E}\left[\exp\left(\lambda m S\right)\right] \\
&= \left[1 + \frac{m}{p-m}\left(\mathrm{e}^{\lambda m} - 1\right)\right]^m \\
&\overset{(a)}{\le} \exp\left(\frac{2m^2}{p}\left(\left(\frac{ep}{m}\right)^b - 1\right)\right) \\
&\overset{(b)}{\le} \exp\left(\frac{2(\log p)^2}{p}\left(\left(\frac{ep}{\log p}\right)^b - 1\right)\right) \\
&\overset{(c)}{\le} \max_{1 \le p \le 512}\left\{\exp\left(\frac{2(\log p)^2}{p}\left(\left(\frac{ep}{\log p}\right)^b - 1\right)\right)\right\} := \tau(b), \qquad (34)
\end{aligned}$$

where (a) follows because $1 + x \le \exp(x)$ for all $x \in \mathbb{R}$ and $m \le p/2$; (b) follows because $m \le \log p$ and $f(x) = \frac{2x^2}{p}\left(\left(\frac{ep}{x}\right)^b - 1\right)$ in non-decreasing in x; (c) follows because $g(x) = \frac{2(\log x)^2}{x}\left[\left(\frac{ex}{\log x}\right)^b - 1\right]$ is non-increasing when $x \ge 512$; $\tau(0+) = 1$ by definition.

In the rest of the proof we shall focus on the intermediate regime: $\log \frac{ep}{m} \le m \le \frac{p}{4}$. Since S dominates H stochastically,

$$\mathbb{E}\left[\exp\left(\lambda H^2\right)\right] \le \mathbb{E}\left[\exp\left(\lambda S^2\right)\right]. \qquad (35)$$

Let $(t_1, \ldots, t_m) \overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{m}{p-m})$ and $T = \sum_{i=1}^m t_i$, which is an independent copy of S. Next we use a decoupling argument to replace S^2 by ST:

$$\begin{aligned}
\left(\mathbb{E}\left[\exp\left(\lambda S^2\right)\right]\right)^2 &= \left(\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^m s_i^2 + \lambda \sum_{i \ne j} s_i s_j\right)\right]\right)^2 \\
&\le \mathbb{E}\left[\exp\left(2\lambda S\right)\right] \mathbb{E}\left[\exp\left(2\lambda \sum_{i \ne j} s_i s_j\right)\right] \qquad (36) \\
&\le \mathbb{E}\left[\exp\left(2\lambda S\right)\right] \mathbb{E}\left[\exp\left(8\lambda \sum_{i \ne j} s_i t_j\right)\right], \qquad (37) \\
&\le \mathbb{E}\left[\exp\left(2\lambda S\right)\right] \mathbb{E}\left[\exp\left(8\lambda ST\right)\right], \qquad (38)
\end{aligned}$$

where (36) is by Cauchy-Schwartz inequality and (37) is a standard decoupling inequality (see, e.g., (Vershynin, 2011, Theorem 1)).

The first expectation on the right-hand side (38) can be easily upper bounded as follows: Since $m \geq \log \frac{\mathrm{e}p}{m}$, we have $\lambda \leq b$. Using the convexity of the exponential function:

$$\exp(ax) - 1 \leq (\mathrm{e}^a - 1)x, \quad x \in [0,1], \tag{39}$$

we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(2\lambda S\right)\right] &\leq \exp\left(\frac{m^2}{p-m}\left(\mathrm{e}^{2\lambda}-1\right)\right) \leq \exp\left(\frac{4\left(\mathrm{e}^{2b}-1\right)m^2\lambda}{bp}\right) \\ &\leq \exp\left(4\left(\mathrm{e}^{2b}-1\right)\frac{m\log\frac{\mathrm{e}p}{m}}{p}\right) \leq \exp\left(4\left(\mathrm{e}^{2b}-1\right)\right), \end{aligned} \tag{40}$$

where the last inequality follows from $\max_{0\leq x\leq 1} x\log\frac{\mathrm{e}}{x} = 1$.

Next we prove that for some function $\tau' : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\tau'(0+) = 1$,

$$\mathbb{E}\left[\exp\left(8\lambda ST\right)\right] \leq \tau'(b), \tag{41}$$

which, in view of (35), (38) and (40), completes the proof of the lemma. We proceed toward this end by truncating on the value of T. First note that

$$\mathbb{E}\left[\exp\left(8\lambda ST\right)\mathbf{1}_{\left\{T>\frac{1}{8\lambda}\right\}}\right] \leq \mathbb{E}\left[\exp\left(8bT\log\frac{\mathrm{e}p}{m}\right)\mathbf{1}_{\left\{T>\frac{1}{8\lambda}\right\}}\right] \tag{42}$$

where the last inequality follows from $S \leq m$ and $\lambda m \leq b\log\frac{\mathrm{e}p}{m}$. It follows from the definition that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(8bT\log\frac{\mathrm{e}p}{m}\right)\mathbf{1}_{\left\{T>\frac{1}{8\lambda}\right\}}\right] \\ &\leq \sum_{t\geq 1/(8\lambda)} \exp\left(8bt\log\frac{\mathrm{e}p}{m}\right)\binom{m}{t}\left(\frac{m}{p-m}\right)^t \\ &\overset{(a)}{\leq} \sum_{t\geq 1/(8\lambda)} \exp\left(8bt\log\frac{\mathrm{e}p}{m} + t\log\frac{\mathrm{e}m}{t} - t\log\frac{p}{2m}\right) \\ &\overset{(b)}{\leq} \sum_{t\geq 1/(8\lambda)} \exp\left(8bt\log\frac{\mathrm{e}p}{m} + t\log\left(8\mathrm{e}b\log\frac{\mathrm{e}p}{m}\right) - t\log\frac{p}{2m}\right) \\ &\overset{(c)}{\leq} \sum_{t\geq 1/(8\lambda)} \exp\left[-t\left(\log 2 - 8b\log(4\mathrm{e}) - \log\left(8\mathrm{e}b\log(4\mathrm{e})\right)\right)\right] \\ &\overset{(d)}{\leq} \sum_{t\geq 1/(8b)} \exp\left[-t\left(\log 2 - 8b\log(4\mathrm{e}) - \log\left(8\mathrm{e}b\log(4\mathrm{e})\right)\right)\right] := \tau''(b) \end{aligned} \tag{43}$$

where (a) follows because $\binom{m}{t} \leq \left(\frac{\mathrm{e}m}{t}\right)^t$ and $m \leq p/2$; (b) follows because $\frac{m}{t} \leq 8m\lambda \leq 8b\log\frac{\mathrm{e}p}{m}$; (c) follows because $m \leq p/4$ and $b \leq 1/(16\mathrm{e})$; (d) follows because $\lambda \leq b$; $\tau''(0+) = 0$ holds because $\log 2 < 8b\log(4\mathrm{e}) + \log\left(8\mathrm{e}b\log(4\mathrm{e})\right)$ for $b \leq 1/(16\mathrm{e})$.

Recall that $m \geq \log\frac{\mathrm{e}p}{m}$. Then $\lambda = b\left(\frac{1}{m}\log\frac{\mathrm{e}p}{m} \wedge \frac{p^2}{m^4}\right) \leq b\left(1 \wedge \frac{p^2}{m^4}\right)$. Hence, we have

$$\frac{m^2\lambda}{p} \leq b\left(\frac{m^2}{p} \wedge \frac{p}{m^2}\right) \leq b \tag{44}$$

By conditioning on T and averaging with respect to S, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(8\lambda ST\right)\mathbf{1}_{\left\{T\le\frac{1}{8\lambda}\right\}}\right] &\le \mathbb{E}\left[\exp\left(\frac{2m^2}{p}\left(\exp(8\lambda T)-1\right)\right)\mathbf{1}_{\left\{T\le\frac{1}{8\lambda}\right\}}\right]\\ &\overset{(a)}{\le} \mathbb{E}\left[\exp\left(\frac{16\mathrm{e}m^2}{p}\lambda T\right)\right]\\ &\overset{(b)}{\le} \exp\left\{\frac{2m^2}{p}\left(\exp\left(\frac{16\mathrm{e}m^2\lambda}{p}\right)-1\right)\right\}\\ &\overset{(c)}{\le} \exp\left\{\frac{32\mathrm{e}^2m^4}{p^2}\lambda\right\} \overset{(d)}{\le} \exp(32\mathrm{e}^2 b),\end{aligned} \tag{45}$$

where (a) follows from $\mathrm{e}^x - 1 \le e^a x$ for $x \in [0, a]$; (b) follows because $T \sim \mathrm{Bin}(m, \frac{m}{p-m})$ and $p \ge 2m$; (c) follows due to (44) and $16\mathrm{e}b \le 1$; (d) follows because $\lambda \le b\frac{p^2}{m^4}$. Assembling (42), (43) and (45), we complete the proof of (41), hence the lemma. ■