Journal of Machine Learning Research 25 (2024) 1-25 Submitted 4/22; Revised 3/24; Published 5/24

Random Subgraph Detection Using Queries

Wasim Huleihel WASIMH@TAUEX.TAU.AC.IL
Department of Electrical Engineering-Systems
Tel Aviv university
Tel Aviv 6997801, Israel

Arya Mazumdar ARYA@UCSD.EDU
Halıcıoğlu Data Science Institute
University of California San Diego
La Jolla, CA 92093, USA

Soumyabrata Pal SOUMYABRATAPAL13@GMAIL.COM
Adobe Research, India
Bangalore, INDIA

Editor: Andrea Montanari

Abstract

The planted densest subgraph detection problem refers to the task of testing whether in a given (random) graph there is a subgraph that is unusually dense. Specifically, we observe an undirected and unweighted graph on n vertices. Under the null hypothesis, the graph is a realization of an Erdős-Rényi graph with edge probability (or, density) q. Under the alternative, there is a subgraph on k vertices with edge probability $p > q$. The statistical as well as the computational barriers of this problem are well-understood for a wide range of the edge parameters p and q. In this paper, we consider a natural variant of the above problem, where one can only observe a relatively small part of the graph using adaptive edge queries. For this model, we determine the number of queries necessary and sufficient (accompanied with a quasi-polynomial optimal algorithm) for detecting the presence of the planted subgraph. We also propose a polynomial-time algorithm which is able to detect the planted subgraph, albeit with more queries compared to the above lower bound. We conjecture that in the leftover regime, no polynomial-time algorithms exist. Our results resolve two open questions posed in the past literature.

Keywords: Random graphs, statistical inference, planted dense subgraph, adaptive probing, queries

1. Introduction

In the planted densest subgraph (PDS) formulation of community detection, the task is to detect the presence of a small subgraph of size k planted in an Erdős-Rényi random graph. This problem has been studied extensively both from the algorithmic and the information-theoretic perspectives Arias-Castro et al. (2014); Butucea and Ingster (2013); Verzelen et al. (2015); Chen and Xu (2016); Montanari (2015); Candogan and Chandrasekaran (2018); Hajek et al. (2016). Nonetheless, the best known algorithms exhibit a peculiar phenomenon: there appears to be a statistical-computational gap between the minimum k at which this task can be solved and the minimum k at which it can be solved in polynomial-time. Tight

statistical-computational bounds for several parameter regimes of the PDS were recently established through average-case reductions from the planted clique conjecture Ma and Wu (2015); Hajek et al. (2015); Brennan et al. (2018). The regimes in which these problems are information-theoretically impossible, statistically possible but computational hard, and admit polynomial-time algorithms appear to have a common structure.

Recently, models of clustering and community detection that allow active querying for pairwise similarities have become quite popular. This includes active learning, as well as data labeling by amateurs via crowdsourcing. Clever implementation of an interactive querying framework can improve the accuracy of clustering and help in inferring labels of large amount of data by issuing only a small number of queries. Queries can be easily implemented, e.g., via captcha. Non-expert workers in crowdsourcing platforms are often not able to label the items directly, however, it is reasonable to assume that they can compare items and judge whether they are similar or not. Understanding the query complexity to recover hidden structures is a fundamental theoretical question with several applications, from community detection to entity resolution Mazumdar and Saha (2017a,b). For example, analyzing query complexity and designing query-based algorithms is relevant to community recovery in social networks, where access to the connections (edges) between individuals may be limited due to privacy concerns; or the network can be very large so only part of the graph can be sampled.

In this paper, we investigate a natural variant of the PDS problem above, where one can only observe a small part of the graph using non-adaptive edge queries. A precise description of our model as well as our main goals are described next.

1.1 Model Formulation and Goal

To present our model, we start by reminding the reader the basic mathematical formulation of the PDS detection 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. Also, let $\mathcal{G}(n,k,p,q)$ with $p>q$ denote the ensemble of random graphs generated as follows:

- Pick k vertices uniformly at random from $[n] \triangleq \{1,2,\ldots,n\}$, and denote the obtained set by $\mathcal{K}$.
- Any two vertices in $\mathcal{K}$ are connected with probability p; all other pairs of vertices are connected with probability q.

In summary, $\mathcal{G}(n,k,p,q)$ is the ensemble of graphs of size n, where a random subgraph $\mathcal{G}(k,p)$ is *planted* in an Erdős-Rényi random graph $\mathcal{G}(n,q)$; this ensemble is known as the PDS model. The vertices in $\mathcal{K}$ form a *community* with higher density than elsewhere. In this paper, we focus on the regime where both edge probabilities p and q are fixed and independent of n. The PDS *detection problem* is defined as follows.

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

$$\mathcal{H}_0 : \mathsf{G}_n \sim \mathcal{G}(n,q) \qquad \text{vs.} \qquad \mathcal{H}_1 : \mathsf{G}_n \sim \mathcal{G}(n,k,p,q). \tag{1}$$

The statistical and computational barriers of the problem in Definition 1 depend on the parameters (n, k, p, q). Roughly speaking, if the planted subgraph size k decreases, or if the "distance" between the densities p and q decrease, the distributions under the null and alternative hypotheses become less distinguishable. The statistical limits (i.e., necessary and sufficient conditions) for detecting planted dense subgraphs, without any constraints on the computational complexity, were established in Arias-Castro et al. (2014); Verzelen et al. (2015). Interestingly, in the same papers it was observed that state-of-the-art low-complexity algorithms are highly suboptimal. This raised the intriguing question of whether those gaps between the amount of data needed by all computationally efficient algorithms and what is needed for statistically optimal algorithms is inherent. According, quite recently Hajek et al. (2015); Brennan et al. (2018, 2019), tight statistical-computational gaps for several parameter regimes of PDS were established through average-case reductions from the planted clique conjecture (see, Conjecture 4 below).

In this work, we consider a variant of the PDS detection problem where one can only inspect a small part of the graph by *non-adaptive edge queries*, defined as follows.

Definition 2 (Oracle/Edge queries) *Consider a graph $\mathsf{G}_n = ([n], \mathcal{E})$ with n vertices, where $\mathcal{E}$ denotes the set of edges. An oracle $\mathcal{O} : [n] \times [n] \to \{0, 1\}$, takes as input a pair of vertices $i, j \in [n] \times [n]$, and if $(i, j) \in \mathcal{E}$, namely, there exists an edges between the chosen vertices, then $\mathcal{O}(i, j) = 1$, otherwise, $\mathcal{O}(i, j) = 0$.*

We consider query mechanisms that evolve dynamically over Q steps/queries in the following form: in step number $\ell \in [\mathsf{Q}]$, the mechanism chooses a pair of vertices $\mathsf{e}_\ell \triangleq (i_\ell, j_\ell)$ and asks the oracle whether these vertices are connected by an edge or not. Generally speaking, either *adaptive* or *non-adaptive* query mechanisms can be considered. In the former, the chosen ℓth pair may depend on the previously chosen pairs $\{\mathsf{e}_i\}_{i<\ell}$, as well as on past responses $\{\mathcal{O}(\mathsf{e}_i)\}_{i<\ell}$. In non-adaptive mechanisms, on the other hand, all queries must be made upfront. In this paper we focus mainly on non-adaptive mechanisms. The query-PDS detection problem is defined as follows.

Problem 1 (Query-PDS detection problem) *Consider the PDS detection problem in Definition 1. There is an oracle $\mathcal{O}$ as defined in Definition 2. Find a set of queries $\mathbb{Q} \subseteq [n] \times [n]$ such that $\mathsf{Q} = |\mathbb{Q}|$, and from the oracle answers it is possible to solve (as defined below) the detection problem in* (1)*. Henceforth, we denote this detection problem by $\mathsf{QPDS}(n, k, p, q, \mathsf{Q})$.*

A detection algorithm $\mathcal{A}_n$ for the problem in Definition 1, makes up to Q non-adaptive edge queries, and based on the query responses is tasked with outputting a decision in $\{0, 1\}$. We define the *risk* of a detection algorithm $\mathcal{A}_n$ as the sum of its Type-I and Type-II errors probabilities, namely,

$$\mathsf{R}(\mathcal{A}_n) = \mathbb{P}_{\mathcal{H}_0}(\mathcal{A}_n(\mathsf{G}_n) = 1) + \mathbb{P}_{\mathcal{H}_1}(\mathcal{A}_n(\mathsf{G}_n) = 0), \tag{2}$$

where $\mathbb{P}_{\mathcal{H}_0}$ and $\mathbb{P}_{\mathcal{H}_1}$ denote the probability distributions under the null and the alternative hypothesis, respectively. If $\mathsf{R}(\mathcal{A}_n) \to 0$ as $n \to \infty$, then we say that $\mathcal{A}_n$ solves the detection problem. Our primary goals in this paper are:

- To characterize the statistical limits of $\mathsf{QPDS}(n, k, p, q, \mathsf{Q})$, namely, to derive necessary and sufficient conditions for when its *statistically impossible* and *statistically possible* to solve the detection problem, ignoring algorithmic computational constraints.
- To devise efficient polynomial-time algorithms for $\mathsf{QPDS}(n, k, p, q, \mathsf{Q})$.

1.2 Related Work and Main Contributions

The problem of finding cliques in an Erdős-Rényi random graph under the same edge query model was considered in Feige et al. (2020). It was shown that under certain limitations on the adaptivity of the considered class of algorithms, any algorithm that makes $\mathsf{Q} = O(n^\alpha)$ adaptive edge queries, with $\alpha < 2$, in ℓ rounds finds cliques of size at most $(2-\epsilon)\log_2 n$ where $\epsilon = \epsilon(\alpha, \ell) > 0$. This lower bound should be contrasted with the fact that current state-of-the-art algorithms that make $\mathsf{Q} = O(n^\alpha)$ queries find a clique of size approximately $(1 + \alpha/2)\log_2 n$. This result was later improved in Alweiss et al. (2020), where the dependency of ϵ on ℓ was removed. Closing the gap between those bounds seems to be a challenging open problem. We also mention Ferber et al. (2015, 2017); Conlon et al. (2018), which study the problems of finding a Hamilton cycle, long paths, and a copy of a fixed target graph, in sparse random graphs under the adaptive edge query model. Another recent line of active research is the analysis of the query complexity in certain clustering tasks, such as, the stochastic block model and community detection Mazumdar and Saha (2017a,b); Vinayak and Hassibi (2016); Hartmann et al. (2016); Anagnostopoulos et al. (2016).

Most closely related to our paper are Rácz and Schiffer (2020); Mardia et al. (2020), where the special case of the planted clique model, where $p = 1$ and $q = 1/2$, under the adaptive edge query model, was considered. For this model, assuming unbounded computational resources, upper and lower bounds on the query complexity for both detecting and recovery were established. Specifically, it was shown in Rácz and Schiffer (2020) that no algorithm that makes at most $\mathsf{Q} = o(n^2/k^2)$ adaptive queries to the adjacency matrix of G_n is likely to solve the detection problem. On the other hand, when $k \geq (2+\epsilon)\log_2 n$, for any $\epsilon > 0$, it was shown in Rácz and Schiffer (2020) that there exists an algorithm (not polynomial time) that solves the detection problem by making at least $\mathsf{Q} = (2+\epsilon)\frac{n^2}{k^2}\log_2^2 n$ adaptive queries. For the recovery task, it was shown in Rácz and Schiffer (2020) that no algorithm that makes at most $\mathsf{Q} = o(n^2/k^2 + n)$ adaptive queries exists, while recovery is possible using $\mathsf{Q} = o(n^2/k^2 \log_2^2 n + n\log_2 n)$ adaptive queries. Note that when the whole graph is shown to the algorithm, namely, $\mathsf{Q} = \binom{n}{2}$, then the above detection upper bound boils down to $k > (2+\epsilon)\log_2 n$, which is folklore and well-known to be tight. On the other hand, the above detection lower bound gives $k = O(1)$, which is loose. Sub-linear time algorithms that find the planted clique in the regime $k = \omega(\sqrt{n \log\log n})$ were proposed in Mardia et al. (2020). Specifically, among other things, it was shown that a simple and efficient algorithm can detect the planted clique using $\mathsf{Q} = O(\frac{n^3}{k^3}\log^3 n)$ non-adaptive queries; conversely using the planted clique conjecture, it was shown that a certain class of non-adaptive algorithms cannot detect the planted clique if $\mathsf{Q} = o(\frac{n^3}{k^3})$, suggesting that in the regime where $\frac{n^2}{k^2} \ll \mathsf{Q} \ll \frac{n^3}{k^3}$ polynomial-time algorithms do not exist.

In this paper, we generalize and strengthen the results of Rácz and Schiffer (2020), resolving two open questions raised in the same paper. First, we consider the more general PDS model which allows for arbitrary edge probabilities. While this might seem as a rather

incremental contribution, it turns out that the lower bounding techniques used in Rácz and Schiffer (2020) are quite weak and result in loose bounds on the query complexity for the PDS model. In a nutshell, the main observation in the proof of the lower bound in Rácz and Schiffer (2020) is that if $\mathsf{Q} \ll n^2/k^2$, then with high probability all edge queries will fall outside the planted clique, no matter how strong/sophisticated the query mechanism is. Therefore, detection would be impossible. While the same bound holds for the PDS model as well, it does not capture the intrinsic dependency of the query complexity on the edge densities p and q. More importantly, as was mentioned above, even for the planted clique problem, the results of Rácz and Schiffer (2020) exhibit a polylog gap between the upper and lower bounds on the query complexity. We close this gap by providing asymptotically tight bounds. Specifically, we show that an algorithm that must makes $\mathsf{Q} = \Omega(\frac{n^2}{k^2\chi^4(p||q)} \log^2 n)$ non-adaptive queries to the adjacency matrix of the graph to be able to detect the planted subgraph, where $\chi^2(p||q)$ is the chi-square distance. On the other hand, we devise a quasi-polynomial-time combinatorial algorithm that detects the planted subgraph with high probability by making $\mathsf{Q} = O(\frac{n^2}{k^2\chi^4(p||q)} \log^2 n)$ non-adaptive queries. For the lower bound, we derive first high probability lower and upper bounds on the number of edge queries $\mathscr{C}$ that fall inside the planted subgraph, associated with the optimal query mechanism. Then, we develop a general information-theoretic lower bound on the risk of any algorithm that is given those Q queries, and essentially observes a subgraph of the PDS model, with a planted signal that is an arbitrary sub-structure of the original planted subgraph on $\mathscr{C}$ edges. We also propose a polynomial-time algorithm which is able to detect the planted subgraph using $\mathsf{Q} = \Omega(\frac{n^3}{k^3\chi^2(p||q)} \log^3 n)$ queries. In the leftover regime, where $\frac{n^2}{k^2} \ll \mathsf{Q} \ll \frac{n^3}{k^3}$ and $k = \omega(\sqrt{n})$, we conjecture that no polynomial-time algorithms exist for detection. Finally, as we discuss later in the paper, the generality of our techniques allows for arbitrary *planting* and *noise* distribution $\mathcal{P}$ and $\mathcal{Q}$, respectively, where the PDS model boils down to $\mathcal{P} = \mathsf{Bern}(p)$ and $\mathcal{Q} = \mathsf{Bern}(q)$.

1.3 Main Results

The following theorem determines (up to a constant factor) the number of queries necessary to solve the query-PDS detection problem. Recall that $\chi^2(p||q)$ and $d_{\mathsf{KL}}(p||q)$ denote the chi-square distance and the Kullback-Leibler (KL) divergence between two $\mathsf{Bern}(p)$ and $\mathsf{Bern}(q)$ random variables, respectively. Note that $\chi^2(p||q) = \frac{(p-q)^2}{q(1-q)}$.

Theorem 3 (Detecting a planted densest subgraph) *Consider the* $\mathsf{QPDS}(n, k, p, q)$ *detection problem, and let* $\epsilon > 0$ *be arbitrary. The following statements hold.*

1. *(Non-adaptive lower-bound) The risk of any algorithm* $\mathcal{A}_n$ *that makes at most* Q *non-adaptive edge queries, is* $\mathsf{R}(\mathcal{A}_n) \geq 1 - o(1)$*, if*

$$\mathsf{Q} < (2 - \epsilon) \cdot \frac{n^2}{k^2\chi^4(p||q)} \log^2 \frac{n}{k}. \tag{3}$$

2. *(Statistical sufficiency) Suppose that* $k \geq (2 + \epsilon_0)\frac{\log n}{d_{\mathsf{KL}}(p||q)}$*, for some* $\epsilon_0 > 0$*. It is possible to detect the presence of a planted densest subgraph, i.e,* $\mathsf{R}(\mathcal{A}_n) \leq o(1)$*, by*

making

$$\mathsf{Q} \geq (2+\epsilon) \cdot \frac{n^2}{k^2 d_{\mathsf{KL}}^2(p||q)} \log^2 \frac{n}{k} \tag{4}$$

queries. Moreover, the queries can be non-adaptive.

3. *(Computational sufficiency) Suppose that* $k = \Omega(\sqrt{n \log n / \chi^2(p||q)})$*. There exists a polynomial-time algorithm* $\mathcal{A}_n$ *that can detect the presence of a planted densest subgraph, i.e,* $\mathsf{R}(\mathcal{A}_n) \leq o(1)$*, by making*

$$\mathsf{Q} = O\left(\frac{n^3}{k^3 \chi^2(p||q)} \log^3 n\right) \tag{5}$$

queries. Moreover, the queries can be non-adaptive.

The proof of Theorem 3 is given in Sections 2 and 3. Let us describe briefly the algorithms achieving the query complexities in the second and third items of Theorem 3. The test achieving the query complexity in the second item of Theorem 3 is the *scan test*. In the first step of this algorithm we subsample a set $\mathcal{S} \subset [n]$ of $\mathsf{M} \in \mathbb{N}$ elements, drawn uniformly at random, and then take all pairwise queries among those elements, resulting in $\mathsf{Q} = \binom{\mathsf{M}}{2}$ non-adaptive queries. Observe that at the end of this first step we, in fact, observe the subgraph $\mathsf{G}_{\mathcal{S}}$ of G_n induced by $\mathcal{S}$. Now, given the responses to those queries, in order to distinguish between hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, we search for the densest subgraph (in the sense of the number of active edges) of a certain size in $\mathsf{G}_{\mathcal{S}}$, and then compare the result to a carefully chosen threshold. The test achieving the query complexity in the third item of Theorem 3, on the other hand, is the *degree test* which, roughly speaking, counts the number of "high degree" vertices (i.e., vertices with degrees exceeding a certain threshold) in a randomly chosen induced subgraph of G_n, and then compare the result to a certain threshold. A very similar variant of this degree test was proposed and analyzed in Mardia et al. (2020), for the planted clique problem. It is clear that the combinatorial scan test is computationally expensive (super-polynomial), while the degree test has a polynomial-time complexity. Interestingly, adaptivity is not needed in order to achieve the statistical barrier. It should be emphasized that the condition $k \geq (2+\epsilon_0)\frac{\log n}{d_{\mathsf{KL}}(p||q)}$ in the second part of Theorem 3 is essential because, otherwise, if $k < (2-\epsilon_0)\frac{\log n}{d_{\mathsf{KL}}(p||q)}$ detection is known to be statistically impossible even if we observe/query the whole graph.

Note that the lower and upper bound for the non-adaptive case in (3) and (4) are tight up to a constant factor; the former depend on p and q through the chi-square distance, while the latter through the KL divergence. As we show in the proof, an alternative condition for the scan test to succeed in detection is

$$\mathsf{Q} \geq (2+\epsilon) \cdot \frac{Cn^2}{k^2 \chi^4(p||q)} \log^2 \frac{n}{k}, \tag{6}$$

for some constant $C \geq 8$; the above condition meets the lower bound in (3) up to the constant factor C. We strongly believe that the source of this negligible gap is due to the lower bound, namely, the $\chi^4(p||q)$ factor in (3) can be in fact replaced with $d_{\mathsf{KL}}^2(p||q)$. It

should be emphasized, however, that in the special case of planted clique where $p = 1$ and $q = 1/2$, we have $\chi^4(p||q) = d^2_{\mathsf{KL}}(p||q) = 1$, and thus our bounds in (3) and (4) are tight and fully characterize the statistical limit of detection. This closes the gap in Rácz and Schiffer (2020).

As can be noticed from the second and third items of Theorem 3, there is a significant gap between the query complexity of the optimal algorithm and that of the computationally efficient one. This observation raises the following intriguing question: *what is the sharp condition on (n, k, p, q, Q) under which the problem admits a computationally efficient test with vanishing risk, and conversely, without which no algorithm can detect the planted dense subgraph reliably in polynomial-time?* The gap observed in our problem is common to many contemporary problems in high-dimensional statistics studied over the last few years. Indeed, recently, there has been a success in developing a rigorous notion of what can and cannot be achieved by efficient algorithms. Specifically, a line of work initiated in Berthet and Rigollet (2013) has aimed to explain these statistical-computational gaps by reducing from conjecturally hard average-case problems in computer science, most notably, the planted clique problem, conjectured to be computationally hard in the regime $k = o(\sqrt{n})$. Accordingly, such reductions from planted clique were established to prove tight statistical-computational gaps for a wide verity of detection and recovery problems, e.g., Berthet and Rigollet (2013); Ma and Wu (2015); Cai et al. (2017); Hajek et al. (2015); Wang et al. (2016a,b); Gao et al. (2017); Brennan et al. (2018, 2019); Wu and Xu (2020); Brennan and Bresler (2020).

As mentioned above, it is widely believed that the planted clique detection problem cannot be solved in randomized polynomial time when $k = o(\sqrt{n})$, which we shall refer to as the *planted clique conjecture*, stated as follows. Below, we let $(\mathcal{H}_0^{\mathsf{PC}}, \mathcal{H}_1^{\mathsf{PC}})$ denote the planted clique detection problem, and we recall that it is a special case of the PDS detection problem with edge probabilities $p = 1$ and $q = 1/2$; for simplicity of notation we designate $\mathsf{PC}(n, k) = \mathsf{PDS}(n, k, 1, 1/2)$.

Conjecture 4 (Planted clique conjecture) *Suppose that $\{\mathcal{A}_n\}$ is a sequence of randomized polynomial-time algorithms $\mathcal{A}_n$ and $\{k_n\}$ is a sequence of positive integers satisfying that $\limsup_{n\to\infty} \frac{\log k_n}{\log n} < 1/2$. Then, if G_n is an instance of $\mathsf{PC}(n, k)$, it holds that*

$$\liminf_{n\to\infty} \left[\mathbb{P}_{\mathcal{H}_0^{\mathsf{PC}}}\left(\mathcal{A}_n(\mathsf{G}_n) = 1\right) + \mathbb{P}_{\mathcal{H}_1^{\mathsf{PC}}}\left(\mathcal{A}_n(\mathsf{G}_n) = 0\right)\right] \geq 1. \tag{7}$$

Going back to our problem, it is clear that for $k = o(\sqrt{n})$, and any $1 \leq \mathsf{Q} \leq \binom{n}{2}$ solving $\mathsf{QPDS}(n, k, \mathsf{Q}, p, q)$ is computationally hard, namely, there exists no randomized polynomial-time *adaptive* algorithm that makes up to Q edge queries and is able to solve the detection problem. Accordingly, Theorem 3 and the planted clique conjecture give a partial phase diagram for when detection is statistically impossible, computational hard, and computational easy. The union of the last two regime is the statistically possible regime. Treating k and Q as polynomials in n, i.e., $\mathsf{Q} = \Theta(n^\alpha)$ and $k = \Theta(n^\beta)$, for some $\alpha \in (0, 2)$ and $\beta \in (0, 1)$, we obtain the phase diagram in Fig. 1. Specifically,

1. *Computationally easy regime (blue region):* there is a polynomial-time algorithm for the detection task when $\alpha > 3 - 3\beta$ and $\beta > 1/2$.

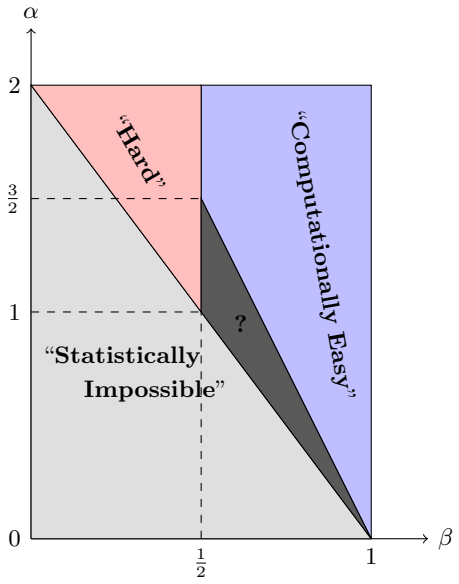

Figure 1: Phase diagram for detecting the presence of a planted dense subgraph, as a function of the dense subgraph size $k = \Theta(n^\beta)$ and the number of non-adaptive edge queries $\mathsf{Q} = \Theta(n^\alpha)$.

2. *Computationally hard regime (red region):* there is an inefficient algorithm for detection when $\beta < 1/2$ and $2 - 2\beta < \alpha$, but the problem is computationally hard (no polynomial-time algorithm exists) in the sense that it is at least as hard as solving the planted clique problem.

3. *Conjecturally hard regime (black region):* there is an inefficient algorithm for detection when $2 - 2\beta < \alpha < 3 - 3\beta$, but we conjecture that there is no polynomial-time algorithm. This was also conjectured in Mardia et al. (2020).

4. *Statistically impossible regime:* the task is statistically/information-theoretically impossible when $\alpha < 2 - 2\beta$.

It turns out that our techniques allows for a more general submatrix detection problem with arbitrary planting and noise distribution $\mathcal{P}$ and $\mathcal{Q}$, respectively, defined as follows.

Definition 5 (General submatrix detection) *Given a pair of distributions $(\mathcal{P}, \mathcal{Q})$ over a measurable space $(\mathcal{X}, \mathcal{B})$, let $\mathsf{SD}(n, k, \mathcal{P}, \mathcal{Q})$ denote the hypothesis testing problem with observation $\mathsf{X} \in \mathcal{X}^{n\times n}$ and hypotheses*

$$\mathcal{H}_0 : \mathsf{X} \sim \mathcal{Q}^{\otimes n\times n} \qquad \text{vs.} \qquad \mathcal{H}_1 : \mathsf{X} \sim \mathcal{D}(n, k, \mathcal{P}, \mathcal{Q}), \tag{8}$$

where $\mathcal{D}(n, k, \mathcal{P}, \mathcal{Q})$ is the distribution of symmetric matrices X with entries $\mathsf{X}_{ij} \sim \mathcal{P}$ if $i, j \in \mathcal{K}$ and $\mathsf{X}_{ij} \sim \mathcal{Q}$ otherwise that are conditionally independent given $\mathcal{K}$, which is chosen uniformly at random over all k-subsets of $[n]$.

Now, consider the following natural generalization of the edge query oracle in Definition 6 to the above submatrix problem.

Definition 6 (Oracle/Entries queries) *Consider an $n \times n$ symmetric matrix $\mathsf{X} \in \mathbb{R}^{n\times n}$. An oracle $\mathcal{O}_{\mathsf{sub}} : [n] \times [n] \to \mathbb{R}$, takes as input a pair of vertices $i, j \in [n] \times [n]$, and outputs $\mathcal{O}_{\mathsf{sub}}(i, j) = \mathsf{X}_{ij}$ in response.*

Then, we define the query-submatrix detection problem as follows.

Problem 2 (Query-submatrix detection problem) *Consider the submatrix detection problem in Definition 5. There is an oracle $\mathcal{O}_{\mathsf{sub}}$ as defined in Definition 6. Find a set of queries $\mathbb{Q} \subseteq [n] \times [n]$ such that $\mathsf{Q} = |\mathbb{Q}|$, and from the oracle answers it is possible to solve the detection problem in* (8)*. Henceforth, we denote this detection problem by $\mathsf{QSD}(n, k, \mathcal{P}, \mathcal{Q}, \mathsf{Q})$.*

It is clear that the PDS model correspond to $\mathcal{P} = \mathsf{Bern}(p)$ and $\mathcal{Q} = \mathsf{Bern}(q)$, while X is the graph adjacency matrix. Then, it can be shown that under mild conditions on the tails of the distribution $\mathcal{P}$ and $\mathcal{Q}$, Theorem 3 holds for the setting in Definition 5 with $\chi^2(p||q)$ and $d_{\mathsf{KL}}(p||q)$ replaced by $\chi^2(\mathcal{P}||\mathcal{Q})$ and $d_{\mathsf{KL}}(\mathcal{P}||\mathcal{Q})$, respectively. Specifically, the lower bound in Theorem 3 holds whenever $0 < \chi^2(\mathcal{P}||\mathcal{Q}) < \infty$, and the upper bounds hold if, for example, the log-likelihood ratio $\mathcal{L} \triangleq \log \frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\mathcal{Q}}$ is sub-Gaussian, or, even slightly weaker; if there is a constant $C \geq 1$ such that

$$\psi_{\mathcal{P}}(\lambda) - d_{\mathsf{KL}}(\mathcal{P}\|\mathcal{Q}) \cdot \lambda \leq C \cdot d_{\mathsf{KL}}(\mathcal{P}\|\mathcal{Q}) \cdot \lambda^2 \quad \forall \lambda \in [-1, 0], \tag{9a}$$

$$\psi_{\mathcal{Q}}(\lambda) + d_{\mathsf{KL}}(\mathcal{Q}\|\mathcal{P}) \cdot \lambda \leq C \cdot d_{\mathsf{KL}}(\mathcal{Q}\|\mathcal{P}) \cdot \lambda^2 \quad \forall \lambda \in [-1, 1], \tag{9b}$$

where $\psi_{\mathcal{Q}} \triangleq \log \mathbb{E}_{\mathcal{Q}}[\exp(\lambda\mathcal{L})]$ and $\psi_{\mathcal{P}} \triangleq \log \mathbb{E}_{\mathcal{P}}[\exp(\lambda\mathcal{L})]$. Note that the above assumption was considered also in Hajek et al. (2017); Brennan et al. (2019), and arise naturally from classical binary hypothesis testing, in order to control the tails of the error probabilities associated with the simple tests we propose.

2. Algorithms and Upper Bounds

In this section we prove items 2 and 3 in Theorem 3. To that end, we propose two algorithms whose performance match (4)–(5). Below, we denote the adjacency matrix of the underlying graph G_n by $\mathsf{A} \in \{0, 1\}^{n\times n}$, with its (i, j) entry denoted by A_{ij}, for any $1 \leq i, j \leq n$.

2.1 Scan Test

In this subsection we analyzed the scan test in Algorithm 1. The parameters M, N_0, and τ_{scan} in Algorithm 1 will be specified below. In the first step of the scan test we subsample a set $\mathcal{S} \subseteq [n]$ of $\mathsf{M} \in \mathbb{N}$ elements, drawn uniformly at random, and take all pairwise queries among these elements. Therefore, the number of queries is $\mathsf{Q} = \binom{\mathsf{M}}{2}$. Given these queries we, in fact, learn the induced subgraph, denoted by $\mathsf{G}_{\mathcal{S}}$, on the set of vertices $\mathcal{S}$. Then, using this subgraph, we wish to distinguish between $\mathcal{H}_0$ and $\mathcal{H}_1$. Recall that $\mathcal{K}$ denotes the set of vertices over which the densest subgraph was planted under $\mathcal{H}_1$, and let N denote the number of planted dense subgraph vertices in $\mathcal{S}$, i.e., $\mathsf{N} \triangleq |\mathcal{K} \cap \mathcal{S}|$. Since $\mathcal{K}$ and $\mathcal{S}$ are drawn

Algorithm 1 `Scan Test`

Require: G_n, $\mathsf{Q} = \binom{\mathsf{M}}{2}$, $\mathsf{N}_0 = (1-\epsilon)\frac{k\mathsf{M}}{n}$, and $\tau_{\mathsf{scan}} = \binom{\mathsf{N}_0}{2} \cdot \gamma$, for $\epsilon \in (0,1)$ and $\gamma \in [q,p]$.

1: Subsample a set $\mathcal{S}$ of M elements drawn uniformly at random from $[n]$.

2: Take all pairwise queries among the elements in $\mathcal{S}$, and obtain A_{ij}, for all $i,j \in \mathcal{S}$.

3: Compute

$$\mathsf{S}_{\mathsf{scan}} \triangleq \max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0} \sum_{i<j\in\mathcal{L}} \mathsf{A}_{ij}.$$

4: If $\mathsf{S}_{\mathsf{scan}} > \tau_{\mathsf{scan}}$ decide $\mathcal{H}_1$; otherwise, decide $\mathcal{H}_0$.

uniformly at random from all sets of size k and M from $[n]$, respectively, we observe that $\mathsf{N} \sim \mathsf{Hypergeometric}(n,k,\mathsf{M})$, namely, N has a Hypergeometric distribution with parameters n, k, and M. Accordingly, we have that $\mathbb{E}(\mathsf{N}) = \frac{k\mathsf{M}}{n}$, and $\mathsf{var}(\mathsf{N}) \leq \frac{k\mathsf{M}}{n}\cdot(1-k/n)$. Therefore, Chebyshev's inequality implies that

$$\mathbb{P}\left[\mathsf{N} \leq (1-\epsilon)\mathbb{E}(\mathsf{N})\right] \leq \frac{1}{\epsilon^2\mathbb{E}(\mathsf{N})}. \tag{10}$$

Thus, provided that $\epsilon^2\mathbb{E}(\mathsf{N}) \to \infty$, we see that with probability tending to unity $\mathsf{N} \geq (1-\epsilon)\frac{k\mathsf{M}}{n} \triangleq \mathsf{N}_0$. The implication of this is the following: as mentioned above we are tasked with the following detection problem:

$$\mathcal{H}_0' : \mathsf{G}_{\mathcal{S}} \sim \mathcal{G}(\mathsf{M},q) \qquad \mathsf{vs.} \qquad \mathcal{H}_1' : \mathsf{G}_{\mathcal{S}} \sim \mathcal{G}(\mathsf{M},\mathsf{N},p,q). \tag{11}$$

Therefore, (11) represents a $\mathsf{PDS}(\mathsf{M},\mathsf{N},p,q)$ detection problem, but the size of the planted densest subgraph is random. Nonetheless, due to (10), it should be clear that by replacing N with N_0 in (11), the detection problem becomes algorithmically harder; thus upper bounds on $\mathsf{PDS}(\mathsf{M},\mathsf{N}_0,p,q)$ imply corresponding upper bounds on $\mathsf{PDS}(\mathsf{M},\mathsf{N},p,q)$. Below, we prove this rigorously.

Recall that A is the adjacency matrix of G_n. The scan test is defined as follows

$$\mathcal{A}_{\mathsf{scan}}(\mathsf{A}_{\mathcal{S}}) \triangleq \mathbb{1}\left\{\max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0} \sum_{i<j\in\mathcal{L}} \mathsf{A}_{ij} \geq \tau_{\mathsf{scan}}\right\}, \tag{12}$$

where $\tau_{\mathsf{scan}} \triangleq \binom{\mathsf{N}_0}{2}\cdot\gamma$, for some $\gamma \in [q,p]$. Under the null hypothesis, for any fixed subset $\mathcal{L}$ of size N_0, it is clear that $\sum_{i<j\in\mathcal{L}} \mathsf{A}_{ij} \sim \mathsf{Binomial}\left(\binom{\mathsf{N}_0}{2}, q\right)$. By the union bound and

Chernoff's inequality,

$$\mathbb{P}_{\mathcal{H}_0'}\left(\mathcal{A}_{\mathsf{scan}}(\mathsf{A}_{\mathcal{S}})=1\right)=\mathbb{P}_{\mathcal{H}_0'}\left[\max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}\geq\tau_{\mathsf{scan}}\right] \tag{13}$$

$$\leq\sum_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\mathbb{P}_{\mathcal{H}_0'}\left[\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}\geq\tau_{\mathsf{scan}}\right] \tag{14}$$

$$\leq\binom{\mathsf{M}}{\mathsf{N}_0}\cdot\mathbb{P}\left[\mathsf{Binomial}\left(\binom{\mathsf{N}_0}{2},q\right)\geq\tau_{\mathsf{scan}}\right] \tag{15}$$

$$\leq\left(\frac{e\mathsf{M}}{\mathsf{N}_0}\right)^{\mathsf{N}_0}\exp\left(-\binom{\mathsf{N}_0}{2}d_{\mathsf{KL}}(\gamma||q)\right) \tag{16}$$

$$=\exp\left(\mathsf{N}_0\log\frac{e\mathsf{M}}{\mathsf{N}_0}-\binom{\mathsf{N}_0}{2}d_{\mathsf{KL}}(\gamma||q)\right). \tag{17}$$

Under the alternative hypothesis, conditioned on $\mathsf{N}=\mathsf{N}'$, for some $\mathsf{N}'\geq\mathsf{N}_0$, the scan test statistics $\max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}$ stochastically dominates $\mathsf{Binomial}\left(\binom{\mathsf{N}_0}{2},p\right)$. By the multiplicative Chernoff's bound,

$$\mathbb{P}_{\mathcal{H}_1'}\left(\mathcal{A}_{\mathsf{scan}}(\mathsf{A}_{\mathcal{S}})=0\right)=\mathbb{P}_{\mathcal{H}_1'}\left[\max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}\leq\tau_{\mathsf{scan}}\right] \tag{18}$$

$$\leq\mathbb{P}\left(\mathsf{N}\leq\mathsf{N}_0\right)+\mathbb{P}\left[\mathsf{Binomial}\left(\binom{\mathsf{N}_0}{2},p\right)\leq\tau_{\mathsf{scan}}\right] \tag{19}$$

$$\leq\frac{1}{\epsilon^2\mathbb{E}\mathsf{N}_0}+\exp\left(-\frac{\binom{\mathsf{N}_0}{2}p}{2}\left(1-\frac{\gamma}{p}\right)^2\right), \tag{20}$$

which clearly goes to 0 since $\mathsf{N}_0\to\infty$. Therefore, combining (17) and (20) we see that $\mathsf{R}(\mathcal{A}_{\mathsf{scan}})\leq\delta$, if

$$d_{\mathsf{KL}}(\gamma||q)>\frac{2}{\mathsf{N}_0}\log\frac{\mathsf{M}}{\mathsf{N}_0}+\frac{2}{\mathsf{N}_0^2}\log\frac{1}{\delta}\geq\frac{\sqrt{2}n}{\sqrt{\mathsf{Q}}k}\log\frac{n}{k}+\frac{n^2}{\mathsf{Q}k^2}\log\frac{1}{\delta}. \tag{21}$$

By taking γ arbitrary close to p, we get the query complexity bound in (4), for arbitrary $\epsilon>0$. Also, note that the condition $k\geq(2+\epsilon_0)\frac{\log n}{d_{\mathsf{KL}}(p||q)}$, for some $\epsilon_0>0$, in the second part of Theorem 3 follows from the constraint that $\mathsf{M}\leq n$. As we mentioned right after Theorem 3, a weaker bound exhibiting a dependency on the chi-square distance can be

derived. Specifically, let $\tau_{\mathsf{scan}} = \binom{\mathsf{N}_0}{2}\frac{p+q}{2}$. By the union bound and Bernstein's inequality,

$$\mathbb{P}_{\mathcal{H}_0'}\left(\mathcal{A}_{\mathsf{scan}}(\mathsf{A}_{\mathcal{S}})=1\right) = \mathbb{P}_{\mathcal{H}_0'}\left[\max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}\geq\tau_{\mathsf{scan}}\right] \tag{22}$$

$$\leq \sum_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\mathbb{P}_{\mathcal{H}_0'}\left[\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}\geq\tau_{\mathsf{scan}}\right] \tag{23}$$

$$\leq \binom{\mathsf{M}}{\mathsf{N}_0}\cdot\mathbb{P}\left[\mathsf{Binomial}\left(\binom{\mathsf{N}_0}{2},q\right)\geq\tau_{\mathsf{scan}}\right] \tag{24}$$

$$\leq \left(\frac{e\mathsf{M}}{\mathsf{N}_0}\right)^{\mathsf{N}_0}\exp\left(-\frac{\binom{\mathsf{N}_0}{2}^2(p-q)^2/4}{2\binom{\mathsf{N}_0}{2}q+\binom{\mathsf{N}_0}{2}(p-q)/3}\right) \tag{25}$$

$$= \exp\left(\mathsf{N}_0\log\frac{e\mathsf{M}}{\mathsf{N}_0}-C\mathsf{N}_0^2\frac{(p-q)^2}{q(1-q)}\right), \tag{26}$$

for some constant $C>8$, and note that $\frac{(p-q)^2}{q(1-q)}=\chi^2(p||q)$. Similarly to (20), under the alternative hypothesis, conditioned on $\mathsf{N}=\mathsf{N}'$, for some $\mathsf{N}'\geq\mathsf{N}_0$, the scan test statistics $\max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}$ stochastically dominates $\mathsf{Binomial}\left(\binom{\mathsf{N}_0}{2},p\right)$. By the multiplicative Chernoff's bound,

$$\mathbb{P}_{\mathcal{H}_1'}\left(\mathcal{A}_{\mathsf{scan}}(\mathsf{A}_{\mathcal{S}})=0\right) = \mathbb{P}_{\mathcal{H}_1'}\left[\max_{\mathcal{L}\subset\mathcal{S}:|\mathcal{L}|=\mathsf{N}_0}\sum_{i<j\in\mathcal{L}}\mathsf{A}_{ij}\leq\tau_{\mathsf{scan}}\right] \tag{27}$$

$$\leq \mathbb{P}\left(\mathsf{N}\leq\mathsf{N}_0\right)+\mathbb{P}\left[\mathsf{Binomial}\left(\binom{\mathsf{N}_0}{2},p\right)\leq\tau_{\mathsf{scan}}\right] \tag{28}$$

$$\leq \frac{1}{\epsilon^2\mathbb{E}\mathsf{N}_0}+\exp\left(-\frac{\left(2\binom{\mathsf{N}_0}{2}-\binom{\mathsf{N}_0}{2}\right)^2(p-q)^2}{8\binom{\mathsf{N}_0}{2}p}\right) \tag{29}$$

$$= \frac{1}{\epsilon^2\mathbb{E}\mathsf{N}_0}+\exp\left(-C\mathsf{N}_0^2q\right), \tag{30}$$

which converges to 0. Therefore, combining (26) and (30) we see that $\mathsf{R}(\mathcal{A}_{\mathsf{scan}})\leq\delta$, if

$$\chi^2(p||q)\geq\Omega\left(\frac{1}{\mathsf{N}_0}\log\frac{\mathsf{M}}{\mathsf{N}_0}+\frac{1}{\mathsf{N}_0^2}\log\frac{2}{\delta}\right)=\Omega\left(\frac{n}{\sqrt{\mathsf{Q}}k}\log\frac{n}{k}+\frac{n^2}{\mathsf{Q}k^2}\log\frac{2}{\delta}\right). \tag{31}$$

2.2 Degree Test

In this subsection we analyze the degree test in Algorithm 2. The parameters n', M, and N_0 in Algorithm 2 will be specified below. Specifically, let $\mathcal{S}$ denote a set of $n'\in\mathbb{N}$ vertices drawn uniformly at random from $[n]$, and $\mathsf{G}_{\mathcal{S}}$ be the subgraph of $\mathcal{G}$ induced by this set. Then, we subsample $\mathsf{M}\in\mathbb{N}$ elements from $\mathcal{S}$, and denote those set of elements by $\mathcal{M}$. Our

Algorithm 2 `Degree Test`

Require: G_n, $n' = \Omega\left(\frac{n^2 \log n}{k^2 \chi^2(p||q)}\right)$, $\mathsf{M} = \Omega\left(\frac{n}{k}\log^2 n\right)$, and $\mathsf{N}_0 = (1-\epsilon)\frac{kn'}{n}$, for $\epsilon \in (0,1)$.

1: Let $\mathcal{S}$ denote a set of n' vertices drawn uniformly at random from $[n]$, and $\mathsf{G}_{\mathcal{S}}$ be the subgraph of $\mathcal{G}$ induced by $\mathcal{S}$.

2: Subsample M elements uniformly at random from $\mathcal{S}$, and denote those set of elements by $\mathcal{M}$.

3: Compute

$$\mathsf{C}_{\mathsf{deg}} \triangleq \sum_{i\in\mathcal{M}} \mathbb{1}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} > n'q + \frac{\mathsf{N}_0(p-q)}{2}\right].$$

4: If $\mathsf{C}_{\mathsf{deg}} > 2\log n'$ decide $\mathcal{H}_1$; otherwise, decide $\mathcal{H}_0$.

test is defined as follows,

$$\mathcal{A}_{\mathsf{deg}}(\mathsf{A}_{\mathcal{S}}) \triangleq \mathbb{1}\left\{\sum_{i\in\mathcal{M}} \mathbb{1}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} > \tau_{\mathsf{deg}}\right] \geq 2\log n'\right\}, \tag{32}$$

where $\tau_{\mathsf{deg}} \triangleq n'q + \frac{\mathsf{N}_0(p-q)}{2}$, and $\mathsf{N}_0 \in \mathbb{N}$. As in the previous subsection, under $\mathcal{H}_1$, let N denote the number of planted dense subgraph vertices in $\mathcal{S}$, i.e., $\mathsf{N} \triangleq |\mathcal{K}\cap\mathcal{S}|$, and note that $\mathsf{N} \sim \mathsf{Hypergeometric}(n,k,n')$. We have shown that with probability tending to unity $\mathsf{N} \geq (1-\epsilon)\frac{kn'}{n} \triangleq \mathsf{N}_0$. Let this event be denoted by $\mathcal{C}_1$, and so $\mathbb{P}(\mathcal{C}_1) \to 1$. Next, we analyze the Type-I+II error probability associated with the above test. Let us start with the null hypothesis. For any $i \in \mathcal{M}$, we let $\mathsf{X}_i \triangleq \mathbb{1}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} > n'q + \frac{\mathsf{N}_0(p-q)}{2}\right] \sim \mathsf{Bern}\left(\mathbb{P}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} > n'q + \frac{\mathsf{N}_0(p-q)}{2}\right]\right)$. Under $\mathcal{H}_0$, it is clear that $\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} \sim \mathsf{Binom}\,(n',q)$, and thus by Bernstein's inequality

$$\mathbb{P}_{\mathcal{H}_0}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} > n'q + \frac{\mathsf{N}_0(p-q)}{2}\right] \leq \exp\left[-\Omega\left(\frac{\mathsf{N}_0^2}{n'}\chi^2(p||q)\right)\right]. \tag{33}$$

Accordingly, if $n' < \frac{\mathsf{N}_0^2\chi^2(p||q)}{C\log n}$, which implies that $n' > \Omega\left(\frac{n^2\log n}{k^2\chi^2(p||q)}\right)$ we will have $\mathbb{P}_{\mathcal{H}_0}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} > n'q + \frac{\mathsf{N}_0(p-q)}{2}\right] \leq n^{-2}$, for some $C > 0$. This implies that under $\mathcal{H}_0$, each X_i is stochastically dominated by $\mathsf{Bern}(n^{-2})$. Consider now the alternative hypothesis. Recall that conditioned on $\mathcal{C}_1$, over the induced subgraph $\mathsf{G}_{\mathcal{S}}$ with n' vertices there are at least N_0 vertices from the planted set. Next, we show that by subsampling M vertices $\mathcal{M}$ from $\mathsf{G}_{\mathcal{S}}$ (as in the second step of Algorithm 2), with high probability among those M samples at least $3\log n'$ vertices fall inside the planted set. To that end, let us divide the subsampled planted vertices $\mathcal{K}\cap\mathcal{S}$ into $3\log n'$ disjoint sets $\{\mathcal{S}_i\}_{i=1}^{3\log n'}$ of equal size $\frac{\mathsf{N}_0}{3\log n'}$. Let $\mathcal{E}$ denote the event that there exist a set $\mathcal{S}_i$ which do not intersect $\mathcal{M}$, namely,

$\mathcal{E} \triangleq \bigcup_{i=1}^{3\log n'} \{\mathcal{S}_i \cap \mathcal{M} = \emptyset\}$. Then, note that

$$\mathbb{P}[\mathcal{E}] \leq (3\log n') \cdot \mathbb{P}[\mathcal{S}_1 \cap \mathcal{M} = \emptyset] \tag{34}$$

$$= (3\log n') \cdot \left[1 - \frac{\mathsf{N}_0}{3n'\log n'}\right]^{\mathsf{M}} \tag{35}$$

$$\leq (3\log n') \cdot \exp\left(-\frac{\mathsf{N}_0\mathsf{M}}{3n'\log n'}\right), \tag{36}$$

namely, the probability that there exists a set $\mathcal{S}_i$ which do not intersect $\mathcal{M}$ is at most (36). Thus, taking $\mathsf{M} \geq \frac{6n'\log^2 n'}{\mathsf{N}_0}$, i.e., $\mathsf{M} = \Omega\left(\frac{n}{k}\log^2 n\right)$ we obtain that $\mathbb{P}[\mathcal{E}] \leq 3\log n'/n'^2$, and accordingly, with probability tending to 1, we sample at least one elements from each $\mathcal{S}_i$. Therefore, among the M samples in $\mathcal{M}$ at least $3\log n'$ vertices fall inside the planted set. Now, for any $i \in \mathcal{M} \cap (\mathcal{K} \cap \mathcal{S})$, we note that $\sum_{j\in\mathcal{S}} \mathsf{A}_{ij}$ stochastically dominates $\mathsf{Binom}(\mathsf{N}_0, p) + \mathsf{Binom}(n - \mathsf{N}_0, q)$, and thus conditioned on $\mathcal{E}^c$ and $\mathcal{C}_1$, by the multiplicative Chernoff's bound,

$$\mathbb{P}_{\mathcal{H}_1}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} \leq n'q + \frac{\mathsf{N}_0(p-q)}{2}\right] \leq \exp\left[-\Omega\left(\frac{\mathsf{N}_0^2}{n'}\chi^2(p||q)\right)\right]. \tag{37}$$

Accordingly, if $n' < \frac{\mathsf{N}_0^2\chi^2(p||q)}{C\log n}$, we will have $\mathbb{P}_{\mathcal{H}_1}\left[\sum_{j\in\mathcal{S}} \mathsf{A}_{ij} \leq n'q + \frac{\mathsf{N}_0(p-q)}{2}\right] \leq n^{-2}$, for large enough $C > 0$. This implies that under $\mathcal{H}_1$, conditioned on $\mathcal{E}^c$ and $\mathcal{C}$ each X_i stochastically dominates $\mathsf{Bern}(1 - n^{-2})$, for $i \in \mathcal{M} \cap (\mathcal{K} \cap \mathcal{S})$. Establishing the above results, we are in a position to upper bound the Type-I and Type-II error probabilities. Using Markov's inequality, we have

$$\mathbb{P}_{\mathcal{H}_0}(\mathcal{A}_{\mathsf{deg}}(\mathsf{A}_{\mathcal{S}}) = 1) = \mathbb{P}_{\mathcal{H}_0}\left[\sum_{i\in\mathcal{M}} \mathsf{X}_i > 2\log n'\right] \tag{38}$$

$$\leq \mathbb{P}\left[\mathsf{Binom}\left(\mathsf{M}, n^{-2}\right) > 2\log n'\right] \tag{39}$$

$$\leq \frac{\mathsf{M}}{2n^2\log n'}, \tag{40}$$

which clearly goes to 0 as $n \to \infty$. On the other hand,

$$\mathbb{P}_{\mathcal{H}_1}(\mathcal{A}_{\mathsf{deg}}(\mathsf{A}_{\mathcal{S}}) = 0) = \mathbb{P}_{\mathcal{H}_1}\left[\sum_{i\in\mathcal{M}} \mathsf{X}_i < 2\log n'\right] \tag{41}$$

$$\leq \mathbb{P}\left[\mathsf{Binom}\left(3\log n', 1 - n^{-2}\right) < 2\log n'\right] + \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{C}^c) \tag{42}$$

$$\leq e^{-C\log n'} + o(1) \to 0, \tag{43}$$

as $n \to \infty$. Finally, note that the number of queries made by Algorithm 2 is $\mathsf{Q} = \mathsf{M} \cdot n' > \Omega\left(\frac{n^3}{k^3\chi^2(p||q)}\log^3 n\right)$ and the constraint that $n' < n$ implies that $k \geq \Omega(\sqrt{n\log n/\chi^2(p||q)})$. This concludes the proof.

3. Statistical Lower Bound

In this section, we prove the lower bounds in Theorem 3, for non-adaptive mechanisms. Our proof consists of two main steps. In the first step, we bound the total number of planted edges any non-adaptive mechanism $\mathscr{Q}_n$ can query. We denote this number of planted queries by $\mathscr{C}(\mathscr{Q}_n)$. Given $\mathscr{Q}_n$, in the second step of the proof, we analyze the resulting detection problem a test $\mathcal{A}_n$ is faced with, and derive a lower bound on its risk. Note that in order to prove that detection is statistically impossible whenever (3) holds, it is suffice to prove impossibility on the boundary, i.e., when $\mathsf{Q} = \mathsf{Q}^\star \triangleq (2-\epsilon)\cdot \frac{n^2}{k^2\chi^4(p||q)} \log^2 n$, for non-adaptive mechanisms. Indeed, if detection is impossible when $\mathsf{Q} = \mathsf{Q}^\star$, then with less queries $\mathsf{Q} < \mathsf{Q}^\star$ detection remains impossible.

3.1 Upper Bound on the Query-Number of Planted Edges

In this subsection, we bound the total number of planted edges any mechanism can query. As mentioned above, this number is denoted by $\mathscr{C}(\mathscr{Q}_n)$ for a given query mechanism $\mathscr{Q}_n$. Letting $\mathbb{Q}$ denote the query trajectory (i.e., the set of Q queried edges), we note that $\mathscr{C}(\mathscr{Q}_n) = \sum_{\ell=1}^{\mathsf{Q}} \mathbb{1}\left\{\mathbb{Q}_\ell \in \mathcal{K}\times\mathcal{K}\right\}$. After making Q queries, a testing algorithm produces an decision. Specifically, given a trajectory $\mathbb{Q}$ (or, a mechanism $\mathscr{Q}_n$), we denote by $(\mathcal{H}_0', \mathcal{H}_1')$ the new hypothesis testing problem faced by a testing procedure: under $\mathcal{H}_0'$ we observe a subgraph over Q with $\mathsf{Bern}(q)$ independent edges, while under $\mathcal{H}_1'$ we observe a subgraph over Q edges, where there exists a set of $\mathscr{C}$ edges that belong to the planted densest subgraph over $\mathcal{K}$ (i.e., $\mathsf{Bern}(p)$ random edges), and the remaining edges are independent $\mathsf{Bern}(q)$ random variables. We denote the respective null and alternative distributions by $\mathbb{P}_{\mathcal{H}_0'}$ and $\mathbb{P}_{\mathcal{H}_1'}$, respectively. We have the following result.

Lemma 7 (Total number of planted edge queries) *Assume that Q satisfies the condition in the first item of Theorem 3, and fix $\delta \in (0,1)$. Let $\mathscr{Q}_n$ be any algorithm that makes at most Q non-adaptive edge queries. Then, with probability at least $1-\delta$,*

$$\mathscr{C}(\mathscr{Q}_n) \le \mathsf{Q}\frac{k^2}{n^2}\left(1 + \frac{1}{\sqrt{\delta}}\frac{n}{k\sqrt{Q}}\right). \tag{44}$$

Remark 8 *Note that on the boundary $\mathsf{Q} = \mathsf{Q}^\star$, we have $\frac{n}{k\sqrt{Q}} = o(1)$, as $n \to \infty$. Therefore, (44) reduce to $\mathscr{C}(\mathscr{Q}_n) \le \mathsf{Q}\frac{k^2}{n^2}(1+o(1))$.*

Proof [Proof of Lemma 7]

We will start by proving Lemma 7 for deterministic query mechanisms, and then, we address the more general case of randomized algorithms. Since queries are made upfront, they are statistically independent of G_n. Accordingly, let X_e, for $e \in \mathbb{Q}$, denote an indicator random variable such that $\mathsf{X}_e = 1$ if $e \in \mathcal{K}\times\mathcal{K}$, and zero otherwise. Then,

$$\mathscr{C}(\mathscr{Q}_n) = \sum_{e\in\mathbb{Q}} \mathsf{X}_e. \tag{45}$$

It is clear that $\mathbb{P}(\mathbb{Q}_\ell \in \mathcal{K}\times\mathcal{K}) = \frac{\binom{k}{2}}{\binom{n}{2}} = \frac{k(k-1)}{n(n-1)} \leq \frac{k^2}{n^2}$, and thus,

$$\mathbb{E}\mathscr{C}(\mathscr{Q}_n) = \frac{\binom{k}{2}}{\binom{n}{2}}\mathsf{Q} = \frac{k(k-1)}{n(n-1)}\mathsf{Q} \triangleq \bar{\mathsf{L}}, \tag{46}$$

and furthermore,

$$\begin{aligned}
\mathbb{E}[\mathscr{C}(\mathscr{Q}_n)]^2 &= \mathbb{E}\sum_{e,e'\in\mathbb{Q}} \mathsf{X}_e\mathsf{X}_{e'} && (47)\\
&= \mathbb{E}\sum_{e\in\mathbb{Q}} \mathsf{X}_e + \mathbb{E}\sum_{e\neq e'\in\mathbb{Q}} \mathsf{X}_e\mathsf{X}_{e'} && (48)\\
&= \frac{k(k-1)}{n(n-1)}\mathsf{Q} + \frac{k(k-1)}{n(n-1)}\sum_{e\neq e'\in\mathbb{Q}} \mathbb{P}\left[e'\in\mathcal{K}\times\mathcal{K}|e\in\mathcal{K}\times\mathcal{K}\right], && (49)
\end{aligned}$$

where the second equality follows the fact that $\mathsf{X}_e^2 = \mathsf{X}_e$, and in the last equality we note that $\mathbb{E}\mathsf{X}_e = \mathbb{P}(e\in\mathcal{K}\times\mathcal{K}) = \frac{\binom{k}{2}}{\binom{n}{2}}$. Fix a pair $e = (i_1^e, j_1^e)$, and consider a pair $e = (i_1^{e'}, j_1^{e'})$. We decompose the summation in (49) into two parts. In the first part, the edges e and e' are completely disjoint, namely, the do not share a common vertex. We denote this by $e\perp e'$. In this case, $\mathbb{P}\left[e'\in\mathcal{K}\times\mathcal{K}|e\in\mathcal{K}\times\mathcal{K}\right] = \frac{\binom{k-2}{2}}{\binom{n-2}{2}} = \frac{(k-2)(k-3)}{(n-2)(n-3)}$. On the other hand, if e and e' share exactly one common vertex, then, $\mathbb{P}\left[e'\in\mathcal{K}\times\mathcal{K}|e\in\mathcal{K}\times\mathcal{K}\right] = \frac{\binom{k-1}{2}}{\binom{n-1}{2}} = \frac{(k-1)(k-2)}{(n-1)(n-2)}$. It is not difficult to show that $\frac{(k-2)(k-3)}{(n-2)(n-3)} \leq \frac{k(k-1)}{n(n-1)}$ and $\frac{(k-1)(k-2)}{(n-1)(n-2)} \leq \frac{k(k-1)}{n(n-1)}$, for $k < n$, and $n = \omega(1)$. Therefore,

$$\sum_{e\neq e'\in\mathbb{Q}} \mathbb{P}\left[e'\in\mathcal{K}\times\mathcal{K}|e\in\mathcal{K}\times\mathcal{K}\right] \leq \frac{k^2(k-1)^2}{n^2(n-1)^2}\mathsf{Q}^2, \tag{50}$$

and thus,

$$\mathbb{E}[\mathscr{C}(\mathscr{Q}_n)]^2 \leq \frac{k(k-1)}{n(n-1)}\mathsf{Q} + \frac{k^2(k-1)^2}{n^2(n-1)^2}\mathsf{Q}^2. \tag{51}$$

Using the above we finally get that,

$$\begin{aligned}
\mathsf{Var}\left(\mathscr{C}(\mathscr{Q}_n)\right) &\leq \frac{k(k-1)}{n(n-1)}\mathsf{Q} + \frac{k^2(k-1)^2}{n^2(n-1)^2}\mathsf{Q}^2 - \frac{k^2(k-1)^2}{n^2(n-1)^2}\mathsf{Q}^2 && (52)\\
&= \frac{k(k-1)}{n(n-1)}\mathsf{Q} = \bar{\mathsf{L}}. && (53)
\end{aligned}$$

Chebyshev's inequality implies that, for any $\epsilon > 0$,

$$\mathbb{P}[\mathscr{C}(\mathscr{Q}_n) \geq (1+\epsilon)\bar{\mathsf{L}}] \leq \frac{1}{\epsilon^2\bar{\mathsf{L}}}. \tag{54}$$

Thus, with probability at least $1-\delta$,

$$\mathscr{C}(\mathscr{Q}_n) \leq \frac{k^2}{n^2}\left(1+\frac{1}{\sqrt{\delta}}\frac{n}{k\sqrt{Q}}\right). \tag{55}$$

Finally, for randomized query mechanisms, we can condition first on the additional source of randomness $\mathcal{R}$, and apply our arguments above for deterministic query mechanisms, to prove (44), independently of the realization of $\mathcal{R}$. Then, by taking an expectation over $\mathcal{R}$ the proof is concluded. ■

3.2 Hypothesis Test Over the Queried Subgraph

Provided with the Q edge queries probed by an arbitrary non-adaptive query mechanism $\mathscr{Q}_n$, we now describe the hypothesis testing problem the detection algorithm $\mathcal{A}_n$ is faced with, and derive a statistical lower bound on its performance. Recall that the overall distinguishing algorithm is a decomposition $\mathcal{A}_n \circ \mathscr{Q}_n$. While we do not specify the distribution of $\mathscr{C}$, we derived in the previous subsection a high probability upper bound on it. Below, we denote the data that is being supplied to the detection algorithm by a vector $\mathbf{X} \in \{0,1\}^{\mathsf{Q}}$ of length Q, with each entry representing a queried edge. While it makes more sense to represent this data using a matrix (which in turn is a sub-matrix of the original adjacency matrix), we find it more convenient to work with the above vector notation.

With some abuse of notation, let $(\mathcal{H}_0', \mathcal{H}_1')$ denote the hypothesis testing problem faced by $\mathcal{A}_n$, followed by the *best* edge query mechanism. Specifically, let $\mathcal{P} \triangleq \mathsf{Bern}(p)$ and $\mathcal{Q} \triangleq \mathsf{Bern}(q)$. Under $\mathcal{H}_0'$ it is clear that $\mathbf{X} \sim \mathbb{P}_{\mathcal{H}_0'}$ where $\mathbb{P}_{\mathcal{H}_0'} \triangleq \mathcal{Q}^{\otimes \mathsf{Q}}$ is the distribution of a product of Q Bernoulli random variables $\mathsf{Bern}(q)$. Under $\mathcal{H}_1'$, the situation is a bit more complicated; we let $\mathbf{X} \sim \mathbb{P}_{\mathcal{H}_1'}$, and the distribution $\mathbb{P}_{\mathcal{H}_1'}$ is defined as follows. The query trajectory $\mathbb{Q}$ completely determines how the query mechanism operates *for any* realization of $\mathcal{K}$. Accordingly, conditioned on $\mathcal{K}$, we let $\mathcal{W}_{\mathcal{K}\times\mathcal{K}}$ denote the set of edge queries that fall inside the planted set $\mathcal{K}\times\mathcal{K}$, associated with the best query mechanism. Then, the alternative distribution is constructed as follows:

1. We pick k vertices uniformly at random from $[n]$, and denote the obtained set by $\mathcal{K}$ (this is the original set of vertices over which the subgraph was planted).
2. Any two vertices in $\mathcal{K}$ are connected with probability p.
3. Let $\mathcal{W}_{\mathcal{K}\times\mathcal{K}} \subset \mathcal{K}\times\mathcal{K}$ be the set of queried edges that fall inside $\mathcal{K}\times\mathcal{K}$, and denote its size by $|\mathcal{W}_{\mathcal{K}\times\mathcal{K}}| = \mathscr{C}$.
4. The elements of $\mathbf{X}$ are the edges in $\mathcal{W}_{\mathcal{K}\times\mathcal{K}}$, and outside this set the edges/elements are drawn i.i.d. $\mathsf{Bern}(q)$.

Note that due to the result in the previous section, such a set $\mathcal{W}_{\mathcal{K}\times\mathcal{K}}$ of random size $\mathscr{C}$ exists, and that the above hypothesis testing problem is *simple*. Below, we let $\mathsf{Unif}_{n,k}$ be the uniform measure over sets of size k drawn u.a.r. from $[n]$. We would like to derive a lower bound on the above testing problem. Specifically, let $\mathsf{R}(\mathcal{A}_n)$ denote the risk of $\mathcal{A}_n$, i.e., the sum of the Type-I and Type-II error probabilities associated with the detection

algorithm $\mathcal{A}_n$. We start by using the following standard lower bound on the risk $\mathsf{R}(\mathcal{A}_n)$ of any algorithm (Tsybakov, 2008, Theorem 2.2),

$$\mathsf{R}(\mathcal{A}_n) \geq 1 - \mathsf{TV}(\mathbb{P}_{\mathcal{H}_0'}, \mathbb{P}_{\mathcal{H}_1'}). \tag{56}$$

Next, recall that in the previous subsection we have proved that $\mathbb{P}(\mathscr{C} > \mathsf{L}^\star) \leq \delta$, where $\mathsf{L}^\star \triangleq \mathsf{Q}\frac{k^2}{n^2}(1+o(1))$. Throughout the rest of the proof, we assume that we are in the regime where $\mathsf{L}^\star \geq 1$ (as it is on the boundary $\mathsf{Q} = \mathsf{Q}^\star$), otherwise, detection is clearly impossible. Indeed, if $\mathsf{L}^\star < 1$, then this implies that the query mechanism do not probe any planted edge. Thus,

$$\begin{aligned}
\mathsf{TV}(\mathbb{P}_{\mathcal{H}_0'}, \mathbb{P}_{\mathcal{H}_1'}) &= \mathsf{TV}(\mathbb{P}_{\mathcal{H}_0'}, \mathbb{E}_{\mathscr{C}}\mathbb{P}_{\mathcal{H}_1'|\mathscr{C}}) && (57)\\
&\leq \mathbb{E}_{\mathscr{C}}\left[\mathsf{TV}(\mathbb{P}_{\mathcal{H}_0'}, \mathbb{P}_{\mathcal{H}_1'|\mathscr{C}})\right] && (58)\\
&\leq \delta + \sum_{\ell \leq \mathsf{L}^\star} \mathsf{TV}(\mathbb{P}_{\mathcal{H}_0'}, \mathbb{P}_{\mathcal{H}_1'|\mathscr{C}=\ell})\mathbb{P}(\mathscr{C} = \ell), && (59)
\end{aligned}$$

where the first inequality follows from the convexity of $(P, Q) \mapsto \mathsf{TV}(P, Q)$. The above total variation distance can be upper bounded as follows (Tsybakov, 2008, Lemma 2.7)

$$2\mathsf{TV}^2(\mathbb{P}_{\mathcal{H}_0'}, \mathbb{P}_{\mathcal{H}_1'|\mathscr{C}=\ell}) \leq \chi^2(\mathbb{P}_{\mathcal{H}_1'|\mathscr{C}=\ell}, \mathbb{P}_{\mathcal{H}_0'}), \tag{60}$$

where $\chi^2(\mathbb{P}_{\mathcal{H}_1'|\mathscr{C}=\ell}, \mathbb{P}_{\mathcal{H}_0'})$ is the chi-square distance between $\mathbb{P}_{\mathcal{H}_1'|\mathscr{C}=\ell}$ and $\mathbb{P}_{\mathcal{H}_0'}$. It should be clear that this total variation is maximized at the boundary, namely, when $\ell = \mathsf{L}^\star$. Accordingly, without loss of generality, below we focus on this case and ignore the conditioning on $\mathscr{C}$, namely, we treat $\mathbb{P}_{\mathcal{H}_1'|\mathscr{C}}$ as $\mathbb{P}_{\mathcal{H}_1'}$ with $\mathscr{C} = \mathsf{L}^\star$.

Let us evaluate the likelihood function $\frac{\mathbb{P}_{\mathcal{H}_1'}}{\mathbb{P}_{\mathcal{H}_0'}}$. Since,

$$\mathbb{P}_{\mathcal{H}_1'} = \mathbb{E}_{\mathcal{K}\sim\mathsf{Unif}_{n,k}}\left[\mathbb{P}_{\mathcal{H}_1'|\mathcal{K},\mathcal{W}_{\mathcal{K}\times\mathcal{K}}}\right], \tag{61}$$

it is clear that,

$$\begin{aligned}
\frac{\mathbb{P}_{\mathcal{H}_1'}}{\mathbb{P}_{\mathcal{H}_0'}}(\mathbf{X}) &= \mathbb{E}_{\mathcal{K}\sim\mathsf{Unif}_{n,k}}\left[\frac{\mathbb{P}_{\mathcal{H}_1'|\mathcal{K},\mathcal{W}_{\mathcal{K}\times\mathcal{K}}}}{\mathcal{Q}^{\otimes\mathsf{Q}}}(\mathbf{X})\right] && (62)\\
&= \mathbb{E}_{\mathcal{K}\sim\mathsf{Unif}_{n,k}}\left[\prod_{i\in\mathcal{W}_{\mathcal{K}\times\mathcal{K}}} f(\mathbf{X}_i)\right], && (63)
\end{aligned}$$

where $f \triangleq \frac{\mathcal{P}}{\mathcal{Q}}$. Thus,

$$\begin{aligned}
\chi^2(\mathbb{P}_{\mathcal{H}_1'}, \mathbb{P}_{\mathcal{H}_0'}) + 1 &= \mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}}\left[\frac{\mathbb{P}_{\mathcal{H}_1'}}{\mathbb{P}_{\mathcal{H}_0'}}(\mathbf{X})\right]^2 && (64)\\
&= \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}}\mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}}\left[\prod_{i\in\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}} f(\mathbf{X}_i)\prod_{i\in\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}} f(\mathbf{X}_i)\right] && (65)
\end{aligned}$$

$$= \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}} \left[\prod_{i\in\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}\cap\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}} f^2(\mathbf{X}_i) \prod_{i\in\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}\Delta\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}} f(\mathbf{X}_i) \right] \tag{66}$$

$$= \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[\prod_{i\in\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}\cap\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}} \mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}} f^2(\mathbf{X}_i) \prod_{i\in\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}\Delta\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}} \mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}} f(\mathbf{X}_i) \right] \tag{67}$$

$$= \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[\prod_{i\in\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}\cap\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}} \mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}} f^2(\mathbf{X}_i) \right] \tag{68}$$

$$= \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[(1+\chi^2(\mathcal{P},\mathcal{Q}))^{|\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}\cap\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}|} \right], \tag{69}$$

where the last equality holds since $\mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}} f(\mathbf{X}_i) = 1$ and $1+\chi^2(\mathcal{P},\mathcal{Q}) = \mathbb{E}_{\mathbb{P}_{\mathcal{H}_0'}} f^2(\mathbf{X}_i)$. To conclude

$$\chi^2(\mathbb{P}_{\mathcal{H}_1'},\mathbb{P}_{\mathcal{H}_0'}) + 1 = \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[(1+\chi^2(\mathcal{P},\mathcal{Q}))^{|\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1}\cap\mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}|} \right]. \tag{70}$$

Now, we note that if $\mathcal{K}_1 \cap \mathcal{K}_2 = \emptyset$ then, obviously, $|\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1} \cap \mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}| = 0$, and thus we may focus on sets $\mathcal{K}_1$ and $\mathcal{K}_2$ such that $|\mathcal{K}_1 \cap \mathcal{K}_2| > 0$. Now, given $\mathcal{K}_1$ and $\mathcal{K}_2$, the random variable $\mathsf{H}_{\mathcal{K}_1\cap\mathcal{K}_2} \triangleq |\mathcal{W}_{\mathcal{K}_1\times\mathcal{K}_1} \cap \mathcal{W}_{\mathcal{K}_2\times\mathcal{K}_2}|$ is clearly upper bounded by $\mathsf{L}^\star \wedge \binom{|\mathcal{K}_1\cap\mathcal{K}_2|}{2}$. Therefore,

$$\chi^2(\mathbb{P}_{\mathcal{H}_1'},\mathbb{P}_{\mathcal{H}_0'}) + 1 \le \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[(1+\chi^2(\mathcal{P},\mathcal{Q}))^{\mathsf{L}^\star\wedge\binom{|\mathcal{K}_1\cap\mathcal{K}_2|}{2}} \right]. \tag{71}$$

Using (60) we have

$$2\cdot\mathsf{TV}(\mathbb{P}_{\mathcal{H}_0'},\mathbb{P}_{\mathcal{H}_1'})^2 \le \chi^2(\mathbb{P}_{\mathcal{H}_1'},\mathbb{P}_{\mathcal{H}_0'}) \tag{72}$$

$$\le \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[(1+\chi^2(\mathcal{P},\mathcal{Q}))^{\mathsf{L}^\star\wedge\binom{|\mathcal{K}_1\cap\mathcal{K}_2|}{2}} \right] - 1 \tag{73}$$

$$\le \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[\exp\left(\mathsf{L}^\star \wedge \binom{|\mathcal{K}_1\cap\mathcal{K}_2|}{2} \cdot \chi^2(\mathcal{P},\mathcal{Q}) \right) \right] - 1. \tag{74}$$

Next, notice that $\mathsf{H} \triangleq |\mathcal{K}_1 \cap \mathcal{K}_2| \sim \mathsf{Hypergeometric}(n,k,k)$, and as so, with this notation, we get

$$\chi^2(\mathbb{P}_{\mathcal{H}_1'},\mathbb{P}_{\mathcal{H}_0'}) \le \mathbb{E}_{\mathcal{K}_1,\mathcal{K}_2\sim\mathsf{Unif}_{n,k}} \left[\exp\left(\mathsf{L}^\star \wedge \binom{\mathsf{H}}{2} \cdot \chi^2(\mathcal{P},\mathcal{Q}) \right) \right] - 1. \tag{75}$$

We analyze the two possible cases: $\mathsf{L}^\star > \binom{\mathsf{H}}{2}$ and $\mathsf{L}^\star \le \binom{\mathsf{H}}{2}$. Furthermore, in the sequel we assume that $k < \sqrt{2n}$, and then discuss the complementary region. Starting with the case where $\mathsf{L}^\star > \binom{\mathsf{H}}{2}$, we have

$$\mathbb{E}\left[\exp\left(\mathsf{L}^\star \wedge \binom{\mathsf{H}}{2} \cdot \chi^2(\mathcal{P},\mathcal{Q}) \right) \mathbb{1}\left\{ \mathsf{L}^\star > \binom{\mathsf{H}}{2} \right\} \right]$$

$$= \mathbb{E}\left[\exp\left(\binom{\mathsf{H}}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q})\right)\mathbb{1}\left\{\mathsf{L}^\star > \binom{\mathsf{H}}{2}\right\}\right]. \tag{76}$$

To analyze this, we use the following tail bound on Hypergeometric random variable (see, e.g., Arias-Castro et al. (2014); Wu and Xu (2023)),

$$\mathbb{P}\left(\mathsf{H}\geq \mathsf{h}\right) \leq \exp\left(-k\cdot \mathsf{d}_{\mathsf{KL}}\left(\mathsf{h}/k||\rho\right)\right), \tag{77}$$

for any $\mathsf{h}/k \geq \rho$, where $\rho \triangleq k/n$. Using the definition of the KL divergence and the identity $(1-x)\log(1-x) \geq -x$, we get

$$\mathsf{d}_{\mathsf{KL}}\left(a||b\right) \geq a\log\frac{a}{b} - a. \tag{78}$$

Using (78), we note that,

$$k\cdot \mathsf{d}_{\mathsf{KL}}\left(\mathsf{h}/k||\rho\right) \geq \mathsf{h}\log\frac{\mathsf{h}}{k\rho} - h \tag{79}$$

$$= \mathsf{h}\log\frac{n\mathsf{h}}{k^2} - h. \tag{80}$$

Accordingly, we have

$$\mathbb{E}\left[\exp\left(\binom{\mathsf{H}}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q})\right)\mathbb{1}\left\{\mathsf{L}^\star > \binom{\mathsf{H}}{2}\right\}\right]$$

$$\leq \mathbb{P}(\mathsf{H}\leq 1) + \sum_{\mathsf{h}=2}^{\sqrt{2\mathsf{L}^\star}} e^{\frac{\mathsf{h}(\mathsf{h}-1)}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q}) - k\cdot\mathsf{d}_{\mathsf{KL}}(\mathsf{h}/k||\rho)} \tag{81}$$

$$\leq 1 + \sum_{\mathsf{h}=2}^{\sqrt{2\mathsf{L}^\star}} e^{\mathsf{h}\left(\frac{\mathsf{h}-1}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q}) - \log\frac{n\mathsf{h}}{k^2}+1\right)}. \tag{82}$$

For $a>0$ fixed, the function $x\to ax-\log x$ is decreasing on $(0,1/a)$ and increasing on $(1/a,\infty)$. Therefore,

$$\frac{\mathsf{h}-1}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q}) - \log\frac{n\mathsf{h}}{k^2} \leq -\omega, \tag{83}$$

where

$$\omega \triangleq \min\left(\log\frac{n}{k^2} - \frac{1}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q}), \log\frac{n\sqrt{2\mathsf{L}^\star}}{k^2} - \frac{\sqrt{2\mathsf{L}^\star}-1}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q})\right). \tag{84}$$

Substituting $\mathsf{L}^\star = k^2\mathsf{Q}/\mathsf{n}^2$, while noticing that $\log\frac{n\sqrt{2\mathsf{L}^\star}}{k^2} = \log\frac{n}{k} + O(\log\frac{\log n}{k}) = [1-o(1)]\cdot\log\frac{n}{k}$, the second term in the minimum tends to ∞ if

$$\mathsf{Q} < (2-\epsilon)\cdot\frac{n^2}{k^2\chi^4(\mathcal{P},\mathcal{Q})}\log^2\frac{n}{k}, \tag{85}$$

for any $\epsilon > 0$. This is also the case of the first term, since

$$\log \frac{n}{k^2} - \frac{1}{2} \cdot \chi^2(\mathcal{P}, \mathcal{Q}) = \log \frac{n\sqrt{\mathsf{L}^\star}}{k^2} - \frac{\sqrt{\mathsf{L}^\star} - 1}{2} \cdot \chi^2(\mathcal{P}, \mathcal{Q}) + \frac{\sqrt{\mathsf{L}^\star}}{2}\chi^2(\mathcal{P}, \mathcal{Q}) - \log \sqrt{\mathsf{L}^\star}, \quad (86)$$

with the second difference bounded from below if $\chi^2(\mathcal{P}, \mathcal{Q})$ is finite. Hence the sum in (82) converges to zero, and we get

$$\mathbb{E}\left[\exp\left(\binom{\mathsf{H}}{2} \cdot \chi^2(\mathcal{P}, \mathcal{Q})\right) \mathbb{1}\left\{\mathsf{L}^\star > \binom{\mathsf{H}}{2}\right\}\right] \leq 1 + o(1). \quad (87)$$

Next, we turn to the case where $\mathsf{L}^\star \leq \binom{\mathsf{H}}{2}$. We have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\mathsf{L}^\star \wedge \binom{\mathsf{H}}{2} \cdot \chi^2(\mathcal{P}, \mathcal{Q})\right) \mathbb{1}\left\{\mathsf{L}^\star \leq \binom{\mathsf{H}}{2}\right\}\right] \\
&\qquad = \exp\left(\mathsf{L}^\star \cdot \chi^2(\mathcal{P}, \mathcal{Q})\right) \cdot \mathbb{P}\left[\mathsf{L}^\star \leq \binom{\mathsf{H}}{2}\right] \quad (88) \\
&\qquad \leq \exp\left(\mathsf{L}^\star \cdot \chi^2(\mathcal{P}, \mathcal{Q}) - \sqrt{2\mathsf{L}^\star} \log \frac{n\sqrt{2\mathsf{L}^\star}}{k^2} + \sqrt{2\mathsf{L}^\star}\right) \quad (89) \\
&\qquad = \exp\left[-\sqrt{2\mathsf{L}^\star}\left(\log \frac{n\sqrt{2\mathsf{L}^\star}}{k^2} - 1 - \sqrt{\frac{\mathsf{L}^\star}{2}} \cdot \chi^2(\mathcal{P}, \mathcal{Q})\right)\right], \quad (90)
\end{aligned}$$

where the inequality follows from (77) and (78). Substituting $\mathsf{L}^\star = k^2\mathsf{Q}/n^2$, and noticing that $\log \frac{n\sqrt{2\mathsf{L}^\star}}{k^2} = \log \frac{n}{k} + O(\log \frac{\log n}{k}) = [1 - o(1)] \cdot \log \frac{n}{k}$, it is clear that the right hand side of (90) converges to zero as long as

$$\mathsf{Q} < (2 - \epsilon) \cdot \frac{n^2}{k^2 \chi^4(\mathcal{P}, \mathcal{Q})} \log^2 \frac{n}{k}, \quad (91)$$

for any $\epsilon > 0$. Accordingly, under this condition (90) converges to zero. Combining (75), (87), and (90), we get that for $k < \sqrt{2n}$,

$$\chi^2(\mathbb{P}_{\mathcal{H}_1'}, \mathbb{P}_{\mathcal{H}_0'}) \leq 1 + o(1) - 1 = o(1), \quad (92)$$

provided that (85) and (91) hold (which coincide), as required (see, (3) in Theorem 3). Finally, we mention that the complementary case, where $k > \sqrt{2n}$ follow from almost the same arguments as above, with the following small modifications; specifically, in this regime, we use the following tail bound on Hypergeometric random variable

$$\mathbb{P}\left(\mathsf{H} < \mathsf{h}\right) \leq \exp\left(-k \cdot \mathsf{d}_{\mathsf{KL}}\left(1 - \mathsf{h}/k || 1 - \rho\right)\right), \quad (93)$$

for any $\mathsf{h}/k < \rho$, where $\rho = k/n$, and we lower bound the KL divergence using,

$$\mathsf{d}_{\mathsf{KL}}\left(a || b\right) \geq -a + (1 - a) \log \frac{1 - a}{1 - b}. \quad (94)$$

Then, repeating the same steps above it can be shown that now the lower tail term is asymptotically small, namely,

$$\mathbb{E}\left[\exp\left(\binom{\mathsf{H}}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q})\right)\mathbb{1}\left\{\mathsf{L}^\star>\binom{\mathsf{H}}{2}\right\}\right]\leq o(1), \tag{95}$$

while the upper tail term is,

$$\mathbb{E}\left[\exp\left(\mathsf{L}^\star\wedge\binom{\mathsf{H}}{2}\cdot\chi^2(\mathcal{P},\mathcal{Q})\right)\mathbb{1}\left\{\mathsf{L}^\star\leq\binom{\mathsf{H}}{2}\right\}\right]\leq 1+o(1), \tag{96}$$

and thus $\chi^2(\mathbb{P}_{\mathcal{H}_1'},\mathbb{P}_{\mathcal{H}_0'})\leq 1+o(1)-1=o(1)$, provided that the same conditions as in (85) and (91) hold. This concludes the proof (see, (3) in Theorem 3).

4. Conclusion and Outlook

In this paper, we formulated and analyzed a variant of the classical PDS problem, where one can only observe a small part of the graph using non-adaptive edge queries. This problem is relevant, for example, when access to the edges (connections) between vertices (individuals) may be scarce due to privacy concerns. For this model, we derived the number of queries necessary and sufficient for detecting the presence of the planted subgraph, up to a constant factor. For the special case of planted cliques, our results are completely tight. This work also has left number of specific problems open, including the following:

- It would be quite interesting to provide any concrete evidence for our conjectured statistical-computational gap either by means of an average-case reduction from the planted clique problem, or failure of classes of powerful algorithms (such as, sum-of-squares hierarchy, low-degree polynomials, etc.), below the computational barrier.
- Our bounds are almost tight in the sense that there is a multiplicative constant gap between our lower and upper bounds. As was mentioned in the paper, we believe that the source for this gap is our lower bound; the $\chi^4(p||q)$ factor should be in fact $d^2_{\mathsf{KL}}(p||q)$. We suspect that one way to prove this is by applying a truncation procedure on the likelihood analysis when deriving the lower bound on the risk.
- In this paper, we analyzed the regime where the edge probabilities p and q are fixed and independent of n. The regime where p and q depend on n, e.g., both decay polynomially in n, is quite important and challenging. It would be interesting to find the phase diagram for this case. Note that here $\chi^2(p||q)$ is not a constant anymore, but decays with n polynomially fast.
- While in this paper we have focused on the detection problem, it is interesting to consider the recovery and partial recovery variants of our setting as well.

Acknowledgments

The authors would like to thank the editor, Andrea Montanari, and the anonymous referee for their suggestions which helped improving the content of this paper. The work of W.

Huleihel was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1734/21). The work of A. Mazumdar was supported by NSF under Award 2217058, and Award 2133484.

References

Ryan Alweiss, Chady Ben Hamida, Xiaoyu He, and Alexander Moreira. On the subgraph query problem. *Combinatorics, Probability and Computing*, page 1–16, Jul 2020.

A. Anagnostopoulos, Jakub Lacki, Silvio Lattanzi, S. Leonardi, and Mohammad Mahdian. Community detection on evolving graphs. In *NIPS*, 2016.

Ery Arias-Castro, Nicolas Verzelen, et al. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 1046–1066, 12–14 Jun 2013.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 648–847, 09–12 Jul 2020.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *COLT*, pages 48–166, 2018.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Universality of computational lower bounds for submatrix detection. In *COLT*, 2019.

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

Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *Annals of Statistics*, 45(4): 1403–1430, 08 2017.

Utkan Onur Candogan and Venkat Chandrasekaran. Finding planted subgraphs with few eigenvalues using the schur–horn relaxation. *SIAM Journal on Optimization*, 28(1):735–759, 2018.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research*, 17(27):1–57, 2016.

D. Conlon, J. Fox, A. Grinshpun, and X. He. Online ramsey numbers and the subgraph query problem. *arXiv: Combinatorics*, pages 159–194, 2018.

Uriel Feige, David Gamarnik, Joe Neeman, Miklós Z. Rácz, and Prasad Tetali. Finding cliques using few probes. *Random Structures & Algorithms*, 56(1):142–153, 2020.

Asaf Ferber, Michael Krivelevich, Benny Sudakov, and Pedro Vieira. Finding hamilton cycles in random graphs with few queries. *Random Struct. Algorithms*, 49, 05 2015. doi: 10.1002/rsa.20679.

Asaf Ferber, M. Krivelevich, B. Sudakov, and Pedro Vieira. Finding paths in sparse random graphs requires many queries. *Random Struct. Algorithms*, 50:71–85, 2017.

Chao Gao, Zongming Ma, and Harrison H Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Information limits for recovering a hidden community. *IEEE Transactions on Information Theory*, 63(8):4729–4745, 2017.

Bruce E Hajek, Yihong Wu, and Jiaming Xu. Computational lower bounds for community detection on random graphs. In *COLT*, pages 899–928, 2015.

Tanja Hartmann, A. Kappes, and D. Wagner. Clustering evolving networks. In *Algorithm Engineering*, 2016.

Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

Jay Mardia, Hilal Asi, and Kabir Aladin Chandrasekher. Finding planted cliques in sub-linear time. *ArXiv*, abs/2004.12002, 2020.

A. Mazumdar and B. Saha. Clustering with noisy queries. In *NIPS*, 2017a.

A. Mazumdar and B. Saha. Query complexity of clustering with side information. In *NIPS*, 2017b.

Andrea Montanari. Finding one community in a sparse graph. *Journal of Statistical Physics*, 161(2):273–299, 2015.

Miklós Z. Rácz and Benjamin Schiffer. Finding a planted clique by adaptive probing. *ALEA Latin American Journal of Probability and Mathematical Statistics*, 17:775–790, 2020.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Nicolas Verzelen, Ery Arias-Castro, et al. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.

Ramya Korlakai Vinayak and B. Hassibi. Crowdsourced clustering: Querying edges vs triangles. In *NIPS*, 2016.

Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of rip certification. In *Advances in Neural Information Processing Systems*, pages 3819–3827, 2016a.

Tengyao Wang, Quentin Berthet, and Richard J Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5): 1896–1930, 2016b.

Yihong Wu and Jiaming Xu. Statistical problems with planted structures: Information-theoretical and computational limits. In Miguel R. D. Rodrigues and Yonina C. Eldar, editors, *Information-Theoretic Methods in Data Science*. Cambridge University Press, Cambridge, 2020.

Yihong Wu and Jiaming Xu. *Statistical inference on graphs: Selected Topics*. Lecture notes, 2023.